Bayesian Transfer Learning and Source Selection

Nicholas G. Polson
Booth School of Business
University of Chicago

Daniel Zantedeschi
Muma College of Business
University of South Florida

August 11, 2026

Abstract

Multi-source transfer learning asks which auxiliary studies should be borrowed from. The Bayesian answer is to place a prior on the informative set and average over it. We study the resulting posterior on source-inclusion configurations and show it is more delicate than the pairwise Bayes factor asymptotics suggest. The informative set admits a predictive definition through partial exchangeability, with a Sanov rate function equal to a design-weighted contrast norm plus a covariate-shift term that the contrast parameterization cannot represent. The Bayes factor for including a source compares two ways of explaining that source's data rather than the source against its absence, so the nuisance specification for excluded sources governs consistency: pooling excluded sources into one coefficient vector gives inclusion probabilities for noninformative sources that do not decay with sample size, while assigning each excluded source its own vector restores consistency once n_k exceeds p. Compatibility is moreover unidentified when the informative pool has one member, since a p-dimensional contrast absorbs any single source's deviation. We establish all three by exact enumeration of the 2^K posterior in a conjugate Gaussian model, which separates the statistical question from Markov chain mixing, and show that integrating out the block variances aggravates rather than repairs the pooled specification, in one case inverting the ordering so that noninformative sources are preferred. Whether any of this reaches the target depends sharply on the target sample size. At the n_0 used in published benchmarks the contrast absorbs the contamination and the choice of nuisance specification moves target error by 0.03%; at an n_0 one third as large, naive pooling produces genuine negative transfer and the same choice is the difference between attaining the oracle and tripling its error. Selection is load-bearing precisely where transfer learning is used rather than benchmarked. We close with a collapsed marginal likelihood for logistic transfer learning under Pólya-Gamma augmentation, a Thorin-class characterization of admissible mixing measures, and a generative computation that never instantiates the configuration vector.

Keywords: Bayes factors; Bayesian model averaging; covariate shift; global-local shrinkage; horseshoe; multiplicity; partial exchangeability; Pólya-Gamma; Sanov's theorem; source selection; sparse contrasts; transfer learning.

1 Introduction

Let $\mathcal{D}_0 = (X^{(0)}, y^{(0)})$ be a target study with n_0 observations on p covariates, $p \gtrsim n_0$, and let $\mathcal{D}_k = (X^{(k)}, y^{(k)})$, $k = 1, \ldots, K$, be auxiliary sources measured on the same covariates. The inferential target is the target regression coefficient $\boldsymbol{\beta}$ alone. Source coefficients are nuisance. Pooling improves efficiency when the sources are close to the target and induces bias when they are not, a failure mode known as negative transfer.

The dominant formalism, due to Li et al. (2022), encodes relatedness through sparse contrasts. Writing $\boldsymbol{\delta}^{(k)} = \boldsymbol{\beta} - \mathbf{w}^{(k)}$, the informative set is

$$\mathcal{A}_q = \big\{1 \le k \le K : \|\boldsymbol{\delta}^{(k)}\|_q \le t\big\}, \tag{1}$$

with t the transferring level. Frequentist procedures estimate $\mathcal{A}$ by screening and then condition on the estimate. Tian and Feng (2023) give confidence intervals valid asymptotically under a fixed empirical $\hat{\mathcal{A}}$. The Bayesian alternative, developed for the multi-source case by Jamshidian (2026), introduces latent inclusion indicators $\boldsymbol{\gamma} \in \{0,1\}^K$ and reports

$$p(\boldsymbol{\beta} \mid \mathcal{D}) = \sum_{\boldsymbol{\gamma} \in \{0,1\}^K} p(\boldsymbol{\beta} \mid \mathcal{D}, \boldsymbol{\gamma})\, p(\boldsymbol{\gamma} \mid \mathcal{D}), \tag{2}$$

so that uncertainty about which sources are informative propagates into the target credible intervals. This is the right thing to do and, to our knowledge, no competing multi-source method does it.

Our concern is with $p(\boldsymbol{\gamma} \mid \mathcal{D})$ itself. The available theory is pairwise. For two configurations $\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}$ assigning at least one study to each block, the models have equal dimension and $\log \mathrm{BF}_{12} = \Delta\ell_n + O_p(1)$ with $n^{-1}\Delta\ell_n \to c > 0$ under the true configuration, giving exponential consistency; at the boundary a $\frac{r}{2}\log n$ penalty appears (Jamshidian, 2026, Thm. 2.4). This is classical and correct. It is also silent about three things that determine whether the posterior in (2) is any good: what the comparison is a comparison of, whether the quantity being compared identifies compatibility at all, and how a search over 2^K configurations behaves at the sample sizes at which transfer learning is actually used.

The empirical symptom is visible in the source material. In simulations with $K = 10$, $p = 200$, $n_k = 150$ and prior inclusion probability $1/2$, truly informative studies attain posterior inclusion probability near 0.7 and noninformative studies near 0.45 (Jamshidian, 2026, Fig. 2.2). For a procedure that is exponentially consistent in theory, that is very weak separation, and it is not explained by Monte Carlo error. We show it is not a mixing failure either.

1.1 Contributions

Our contributions are six, three structural and three empirical. We state them as claims about the framework rather than about any one implementation of it.

(i) *The informative set has a predictive definition, and the parametric one is design dependent.* Definition (1) is a statement about coefficients and so depends on the parameterization and on the scaling of X. We define $\mathcal{A}$ instead as the block of the finest partially exchangeable partition containing the target (Definition 1) and compute the Sanov rate $I_k = (2\sigma^2)^{-1}(\boldsymbol{\delta}^{(k)})^\top \Sigma_k \boldsymbol{\delta}^{(k)} + \mathrm{KL}(P_{x^{(k)}} \| P_{x^{(0)}})$ (Proposition 2). The two definitions agree only up to the conditioning of Σ (Proposition 3), and the covariate-shift term is invisible to the contrast parameterization entirely.

(ii) *The nuisance block, not the contrast, governs selection consistency.* Excluded sources are modelled rather than discarded, so the Bayes factor for γ_k compares two explanations of $y^{(k)}$. Pooling excluded sources into one coefficient vector makes that comparison inherit a misspecification rate that does not vanish, and selection is inconsistent for $\mathcal{A}$ (Proposition 4); giving each excluded source its own vector restores consistency once $n_k/p \to \infty$ (Proposition 5).

(iii) *Compatibility is unidentified at a singleton pool, and the evidence is exactly quantifiable.* A p-dimensional contrast absorbs any single source's deviation, so with a diffuse contrast prior the target data carry no information about γ_k when $|\mathcal{A}_\gamma| = 1$ (Proposition 6). Evidence arises only from mutual coherence among two or more included sources. Under orthogonal designs the flip

statistic is exactly $-2\Delta\ell_k \sim \chi^2_p(2n_k I_k)$ (Proposition 7), which separates the source-independent Occam term from the compatibility signal, yields the sample size a search over 2^K configurations requires, and supplies an unbiased estimator of I_k with a closed-form standard error (Section 4.6).

(iv) *Exact enumeration.* Every empirical claim about these posteriors in the literature is filtered through a Metropolis-within-Gibbs sampler with a tempering schedule, so a weak posterior and a badly mixing chain are not distinguishable from the reported output. We enumerate all 2^K configurations in closed form. This is what lets us attribute the observed weak separation to the posterior rather than to the algorithm, and it is what makes the remaining claims testable at all.

(v) *Freeing the residual variances aggravates the problem.* A misfitting pooled nuisance block has a second way out: inflate its variance rather than relocate the source. Integrating the block variances out analytically, we find this does not repair the specification. At $n_k/p = 1.5$ the posterior excludes every source and no transfer occurs; at $n_k/p = 3$ the ordering inverts, and noninformative sources receive higher inclusion probability than informative ones (Section 5.1).

(vi) *Whether any of it matters is governed by an absorption identity.* In the standard benchmark designs the specification is nearly irrelevant: it moves target error by 0.03% and credible interval widths by 0.17%. Proposition 9 explains why, and gives the exact bias for arbitrary designs, matching the simulated excess to 2%. Its isotropic form $\{|\mathcal{A}|(1 + n_0 g_\delta)\}^{-1}$ identifies $n_0 g_\delta$ as the controlling quantity, so that tightening the contrast prior for identification simultaneously increases exposure to bad sources. Reducing n_0 by a factor of three reverses every conclusion: naive pooling becomes worse than ignoring the sources, and the nuisance specification becomes the difference between attaining the oracle and tripling its error (Section 6.3).

Items (ii), (iii) and (vi) are the ones we would defend hardest. Item (v) rests on a simplified variance hierarchy and three replicates, and we flag its limits where we state it.

1.2 Related work and outline

The sparse contrast formalism is due to Li et al. (2022), who give a two-stage Lasso procedure with an oracle version assuming $\mathcal{A}$ known and an aggregation scheme when it is not. Tian and Feng (2023) extend this to generalized linear models and supply asymptotically valid confidence intervals conditional on an empirically determined informative set. The Bayesian literature is younger. Single-source normal means problems have been treated with horseshoe priors on the contrast; multi-source integration has been approached by centering a shrinkage prior at a weighted average of pre-estimated source coefficients, and by conditional spike-and-slab priors solved variationally. Jamshidian (2026) is, to our knowledge, the first multi-source framework to propagate selection uncertainty into target inference rather than conditioning on a point estimate of $\mathcal{A}$, and it is the object of our analysis throughout.

The tools we bring are standard elsewhere. Global-local shrinkage in the sense of Polson and Scott (2011) and Bhadra et al. (2016), with the horseshoe of Carvalho et al. (2010) as the reference case; automatic multiplicity adjustment through a random inclusion probability, following Scott and Berger (2010); Pólya-Gamma augmentation (Polson et al., 2013) for the binary case; and posterior contraction results for scale mixture priors in high-dimensional regression (Song and Liang, 2023), on which the oracle theory in the source material rests. What is new is the application of these to the selection step rather than to the estimation step, and the observation that the selection step has different failure modes.

Section 2 fixes notation. Section 3 gives the predictive definition of the informative set and the Sanov rate. Section 4 contains the results on nuisance specification, identification, and the sample size required by a search over 2^K configurations. Section 5 reports the enumeration experiments. Section 6 asks

whether any of it affects target inference and answers mostly no. Sections 7 through 9 treat binary and survival outcomes, the admissible class of mixing measures, and computation. Section 10 states what we have not shown. Proofs are collected in the appendix.

2 Setup

2.1 The contrast hierarchy

Conditional on $\boldsymbol{\gamma}$, write $\mathcal{A}_\gamma = \{k : \gamma_k = 1\}$ and $\bar{\mathcal{A}}_\gamma = \{k : \gamma_k = 0\}$. The Gaussian working model is

$$y^{(\mathcal{A}_\gamma)} \sim \mathcal{N}\big(X^{(\mathcal{A}_\gamma)}\mathbf{w}, \sigma^2_{(\mathcal{A})} I\big), \tag{3}$$

$$y^{(0)} \sim \mathcal{N}\big(X^{(0)}(\mathbf{w} + \boldsymbol{\delta}), \sigma^2_{(0)} I\big), \tag{4}$$

$$y^{(\bar{\mathcal{A}}_\gamma)} \sim \mathcal{N}\big(X^{(\bar{\mathcal{A}}_\gamma)}\mathbf{v}, \sigma^2_{(\bar{\mathcal{A}})} I\big), \tag{5}$$

where $\mathbf{w}$ anchors the informative pool, $\boldsymbol{\delta}$ is the target contrast and $\boldsymbol{\beta} = \mathbf{w} + \boldsymbol{\delta}$. All three blocks receive independent normal scale mixtures, $w_j \mid \sigma^2, \nu_j \sim \mathcal{N}(0, \sigma^2\nu_j)$, in the sense of Polson and Scott (2011) and Bhadra et al. (2016), with the horseshoe $\nu_j = \lambda_j^2\tau^2$, $\lambda_j \sim C^+(0,1)$, $\tau \sim C^+(0,\psi^2)$ as the reference implementation (Carvalho et al., 2010). Finally $\gamma_k \mid \pi \sim \mathrm{Bern}(\pi)$.

Two features of (3)–(5) do the work in what follows, and neither is incidental. The first is that $\boldsymbol{\delta}$ is a single p-vector shared by the whole informative pool, so that compatibility is a statement about the pool and not about individual sources. The second is that $\bar{\mathcal{A}}_\gamma$ is *modeled*, through $\mathbf{v}$, rather than discarded. Excluded data are still explained. We will show in Section 4 that this second choice, made for reasons of coherence, is what governs selection consistency.

2.2 The truncation as a bound on the contrast scale

Jamshidian (2026) imposes $\tau_{(0)} \mid \tau_{(\mathcal{A})} \sim C^+(0,\psi^2)\,\mathbb{I}(\tau_{(0)} < \tau_{(\mathcal{A})})$, requiring the contrast to be strictly sparser than the anchor. Read as a modelling convenience this looks arbitrary. It is in fact a bound on how much the contrast is permitted to absorb, and it is worth making the bound explicit, because Proposition 6 says identification depends on it and Proposition 9 says robustness depends on it in the opposite direction.

Under the horseshoe hierarchy the effective prior variance of a coefficient is $\sigma^2\lambda_j^2\tau^2$, so the truncation constrains the ratio of contrast to anchor scale. Write $T = \tau_{(0)}/\tau_{(\mathcal{A})}$. For independent half-Cauchy marginals the ratio has density

$$g(t) = \frac{4}{\pi^2}\,\frac{\log t}{t^2 - 1}, \qquad t > 0, \tag{6}$$

which is positive throughout, integrates to one, and is symmetric under $t \mapsto 1/t$ in the sense that $g(1/t) = t^2 g(t)$. Two moments of the truncated law follow in closed form. Since $\int_0^1 g = \frac{4}{\pi^2}\int_0^1 \log t/(t^2-1)\,dt = \frac{4}{\pi^2}\cdot\frac{\pi^2}{8} = \frac{1}{2}$, the truncation retains exactly half the prior mass, and substituting $s = t^2$,

$$\mathbb{E}\big[T \mid T < 1\big] = \frac{1}{3}, \qquad \mathbb{E}\big[T^2 \mid T < 1\big] = 1 - \frac{8}{\pi^2} \approx 0.189. \tag{7}$$

So the truncation does not merely order the two scales. It places the contrast scale at one third of the anchor scale in expectation, and the contrast *variance* at just under a fifth of the anchor variance. In the conjugate analogue used for the enumeration in Section 5 this corresponds to $g_\delta \approx g_w/5$, and our choice

$g_\delta = 0.1$ against $g_w = 1$ is slightly tighter than the truncation implies, which if anything understates the effects we report in Section 6.3.

The shrinkage factor makes the consequence concrete. For a single coordinate with n effective observations, the posterior mean of a coefficient with prior variance $\sigma^2 g$ is shrunk by $\kappa = (1+ng)^{-1}$ toward zero, so the truncation raises the contrast shrinkage factor from $\kappa_w = (1+n_0 g_w)^{-1}$ to $\kappa_\delta = (1+n_0 g_\delta)^{-1}$, with $\kappa_\delta/\kappa_w \approx 1 + n_0(g_w - g_\delta)/(1+n_0 g_w)$. The contrast is held back by a factor that grows with n_0. This is the quantity that reappears as the absorption factor (25) in Section 6.3, and it is why the identification benefit of the truncation and its robustness cost are governed by the same number, $n_0 g_\delta$.

3 A predictive definition of the informative set

Definition (1) is a statement about parameters. It therefore depends on the parameterization, on the scaling of the design, and on a threshold t with no operational meaning. We prefer a definition in terms of observables.

Let $Z^{(k)} = \{(x_i^{(k)}, y_i^{(k)})\}_{i \le n_k}$ denote the data from study k, $k = 0, 1, \dots, K$. A partition $\mathcal{P}$ of $\{0, 1, \dots, K\}$ induces partial exchangeability if, for every block $B \in \mathcal{P}$, the pooled sequence $\bigcup_{k \in B} Z^{(k)}$ is exchangeable. By de Finetti's representation for partially exchangeable arrays, the observables in block B are conditionally i.i.d. given a block parameter θ_B, with a mixing measure over blocks.

Definition 1 (Predictive informative set)**.** The informative set is the block of the finest partition $\mathcal{P}^\star$ under which partial exchangeability holds that contains the target index 0: $\mathcal{A}^\star = \{k \ge 1 : k \sim_{\mathcal{P}^\star} 0\}$.

This says exactly what transfer learning wants it to say: source k is informative if its observations could have been generated by the same predictive law as the target's. It refers only to observables, so it is invariant under reparameterization and under rescaling of X.

Definition 1 is exact but not quantitative. The rate at which data distinguish blocks is a relative entropy, by Sanov's theorem applied to the empirical measures of the studies.

Proposition 2 (Rate function)**.** *Let P_k denote the joint law of $(x^{(k)}, y^{(k)})$ and P_0 that of the target. Under the Gaussian model (4) with common noise variance σ^2 and covariate second-moment matrix $\Sigma_k = \mathbb{E}[x^{(k)}(x^{(k)})^\top]$,*

$$I_k \;:=\; \mathrm{KL}\big(P_k \,\|\, P_0\big) \;=\; \frac{1}{2\sigma^2}\,(\boldsymbol{\delta}^{(k)})^\top \Sigma_k\, \boldsymbol{\delta}^{(k)} \;+\; \mathrm{KL}\big(P_{x^{(k)}} \| P_{x^{(0)}}\big), \tag{8}$$

where $\boldsymbol{\delta}^{(k)} = \boldsymbol{\beta} - \mathbf{w}^{(k)}$.

Proof. Conditioning on x, the two conditionals are Gaussian with common variance and means differing by $x^\top \boldsymbol{\delta}^{(k)}$, giving $\mathrm{KL} = (x^\top \boldsymbol{\delta}^{(k)})^2/(2\sigma^2)$ pointwise; take expectations under $P_{x^{(k)}}$ and add the marginal covariate term by the chain rule. □

Proposition 3 (When the two definitions agree)**.** *Suppose the covariate laws are common across studies, $P_{x^{(k)}} = P_{x^{(0)}}$ for all k, with $\Sigma_k = \Sigma$ of full rank. Then $\mathcal{A}^\star$ of Definition 1 coincides with the level set $\{k : I_k = 0\}$, and for any $t > 0$ the set $\mathcal{A}_2$ of (1) satisfies*

$$\Big\{k : I_k \le \tfrac{\lambda_{\min}(\Sigma)}{2\sigma^2} t^2\Big\} \;\subseteq\; \mathcal{A}_2 \;\subseteq\; \Big\{k : I_k \le \tfrac{\lambda_{\max}(\Sigma)}{2\sigma^2} t^2\Big\}. \tag{9}$$

The two definitions therefore agree up to the conditioning number of the design, and disagree by an unbounded amount when the design is ill-conditioned, which is the high-dimensional case of interest.

The proof is immediate from (8) and the Rayleigh bounds. The practical content is that a transferring level t calibrated on one design does not transfer to another, and that in the genomic applications where the columns of X are strongly correlated the gap between the two sets can be large.

Three consequences. First, (1) with $q = 2$ is the special case $\Sigma_k = I$ with the covariate term suppressed, so the transferring level is a design-free approximation to a design-weighted quantity. Second, the second term in (8) is covariate shift, which the contrast parameterization has no way to represent: a source with $\boldsymbol{\delta}^{(k)} = 0$ but a different covariate law is counted as fully informative by (1) and as non-exchangeable by Definition 1. This is the gap flagged as open in the source material and it is where transfer learning meets the optimal transport formulation of covariate shift. Third, I_k is the per-observation exponential rate, so the evidence accumulated about γ_k is $n_k I_k$ to first order, which is the quantity we now examine.

4 What the Bayes factor for source inclusion compares

4.1 The nuisance block is not a spectator

Under (3)–(5) the data $y^{(k)}$ appear in the likelihood whether or not $\gamma_k = 1$. The Bayes factor is therefore

$$\log \mathrm{BF}(\gamma_k = 1 : \gamma_k = 0) = \underbrace{\log \frac{m\big(y^{(k)}, \mathcal{D}_0, \mathcal{D}_{\mathcal{A}\setminus k} \mid \text{pooled with } \mathbf{w}\big)}{m\big(\mathcal{D}_0, \mathcal{D}_{\mathcal{A}\setminus k}\big)}}_{\text{fit under the informative block}} - \underbrace{\log \frac{m\big(y^{(k)}, \mathcal{D}_{\bar{\mathcal{A}}} \mid \text{pooled with } \mathbf{v}\big)}{m\big(\mathcal{D}_{\bar{\mathcal{A}}}\big)}}_{\text{fit under the nuisance block}}. \tag{10}$$

The second term is a property of the *other excluded sources*, not of source k's relationship to the target. If the nuisance block is correctly specified, that term is the marginal likelihood of $y^{(k)}$ under a free p-dimensional coefficient and the comparison reduces to $n_k I_k$ plus an estimation cost. If it is misspecified, the comparison is contaminated by a rate that does not vanish.

Proposition 4 (Inconsistency under a pooled nuisance block)**.** *Suppose the excluded sources share a single coefficient vector* $\mathbf{v}$*, as in* (5)*, and suppose the true source coefficients* $\{\mathbf{w}^{(k)}\}_{k \notin \mathcal{A}}$ *are not all equal. Let* $\mathbf{v}^\star(\bar{\mathcal{A}}_\gamma)$ *minimize the aggregate Kullback-Leibler divergence over the excluded block. Then, with* $n_k = n$ *for all* k *and* $n \to \infty$ *with* p *fixed,*

$$\frac{1}{n} \log \mathrm{BF}(\gamma_k = 1 : \gamma_k = 0) \;\longrightarrow\; J_k^{\mathrm{out}}(\bar{\mathcal{A}}_\gamma) - I_k^{\mathrm{in}}(\mathcal{A}_\gamma), \tag{11}$$

where $J_k^{\mathrm{out}}(\bar{\mathcal{A}}_\gamma) = \frac{1}{2\sigma^2}(\mathbf{w}^{(k)} - \mathbf{v}^\star)^\top \Sigma (\mathbf{w}^{(k)} - \mathbf{v}^\star) > 0$ *is the misspecification rate incurred by forcing source* k *into the excluded pool. Both limits are strictly positive and neither vanishes with* n*, so the sign of the limit is a function of the composition of* $\bar{\mathcal{A}}_\gamma$ *rather than of* I_k*. Selection is not consistent for* $\mathcal{A}$*.*

The mechanism is worth stating in words. If the excluded sources are mutually heterogeneous, the excluded pool fits none of them, and the model can reduce total misspecification by relocating some of them into the informative pool, where they are also fitted badly but where the cost is shared with the target and the anchor. The partition that maximizes the marginal likelihood then balances two pooling costs. It is not the partition that identifies compatibility with the target. Increasing n_k does not help, because both rates scale in n_k together.

The remedy is immediate: give each excluded source its own coefficient vector, $y^{(k)} \sim \mathcal{N}(X^{(k)}\mathbf{v}_k, \sigma^2 I)$ for $k \in \bar{\mathcal{A}}_\gamma$. The excluded model is then correctly specified, $J_k^{\mathrm{out}} = 0$, the marginal likelihood factorizes as

$$\log m(y \mid \boldsymbol{\gamma}) \;=\; f(\mathcal{A}_\gamma) \;+\; \sum_{k \notin \mathcal{A}_\gamma} c_k, \tag{12}$$

and the c_k are computed once for all 2^K configurations. The computational saving is incidental; the point is that (12) makes the comparison a genuine one.

Proposition 5 (Consistency under source-specific nuisance)**.** *Under the source-specific nuisance specification, with $\boldsymbol{\delta}$ given a proper prior with positive density at $\boldsymbol{\delta}^\star$, and with $n_k/p \to \infty$,*

$$\frac{1}{n_k} \log \mathrm{BF}(\gamma_k = 1 : \gamma_k = 0) \;\longrightarrow\; -I_k + \frac{1}{2}\frac{p}{n_k} \log \frac{n_k}{2\pi}\big(1 + o(1)\big), \tag{13}$$

so the Bayes factor favours inclusion when the estimation saving from pooling, of order $p \log n_k/(2n_k)$ per observation, exceeds the incompatibility rate I_k.

The second term is the reason $n_k \gtrsim p$ matters. When $n_k < p$ an individual source is not identified on its own, the estimation saving from pooling is of the same order as the misspecification cost, and every source looks worth including regardless of I_k. Section 5 shows this is not a small effect.

4.2 Non-identification with a single informative source

The second structural result concerns $|\mathcal{A}_\gamma| = 1$.

Proposition 6 (Non-identification at $|\mathcal{A}_\gamma| = 1$)**.** *Let $\mathcal{A}_\gamma = \{k\}$ and let $\boldsymbol{\delta}$ have prior $\mathcal{N}(0, \sigma^2 g_\delta I_p)$. As $g_\delta \to \infty$ with $X^{(0)}$ of full row rank, the target contribution to $m(y \mid \boldsymbol{\gamma})$ converges to a quantity free of* $\mathbf{w}$*, and consequently*

$$\log \mathrm{BF}(\gamma_k = 1 : \gamma_k = 0) \;\longrightarrow\; \textit{a function of } \mathcal{D}_k \textit{ alone}. \tag{14}$$

The evidence for the compatibility of source k with the target vanishes.

Sketch. With $\mathcal{A}_\gamma = \{k\}$, the target enters only through $y^{(0)} \sim \mathcal{N}(X^{(0)}(\mathbf{w} + \boldsymbol{\delta}), \sigma^2 I)$. Reparameterize $\boldsymbol{\beta} = \mathbf{w} + \boldsymbol{\delta}$. As $g_\delta \to \infty$ the prior on $\boldsymbol{\beta}$ given $\mathbf{w}$ becomes flat, so the target marginal likelihood factorizes from $\mathbf{w}$ up to a constant. What remains depends on $\mathcal{D}_k$ only. □

The interpretation is that a p-dimensional contrast can absorb any single source's deviation from the target, so the target data carry no information about whether that deviation is small. Evidence about compatibility arises only from *mutual* coherence: with two or more sources in the pool, a single $\mathbf{w}$ must serve both, and disagreement between them is what the marginal likelihood sees. Compatibility with the target is inferred indirectly, through the requirement that the shared anchor also supports the target after a sparse correction.

This is why the truncation $\tau_{(0)} < \tau_{(\mathcal{A})}$ is not decoration. Requiring the contrast to be sparser than the anchor is precisely the constraint that prevents $\boldsymbol{\delta}$ from absorbing everything, and it restores identification at $|\mathcal{A}_\gamma| = 1$ by limiting g_δ. Figure 2 quantifies how much identification it buys.

4.3 The exact law of the selection statistic

Pairwise consistency does not control a search over 2^K candidate partitions, and to say anything about that search we need the sampling distribution of a single flip rather than only its mean. Under orthogonal designs it is available exactly.

Proposition 7 (Law of the flip statistic)**.** *Assume $X^{(k)\top} X^{(k)} = n_k I_p$ with $n_k \geq p$, known σ^2, and the source-specific nuisance specification with prior scale g_v. Let $\Delta\ell_k$ denote the maximized log-likelihood*

difference between the models with $\gamma_k = 1$ and $\gamma_k = 0$, the anchor being estimated with negligible error. Then

$$-2\Delta\ell_k \;\sim\; \chi^2_p(\lambda_k), \qquad \lambda_k = \frac{n_k\|\boldsymbol{\delta}^{(k)}\|_2^2}{\sigma^2} = 2n_k I_k, \tag{15}$$

a noncentral chi-square on p degrees of freedom, and consequently

$$\log \mathrm{BF}(\gamma_k = 1 : \gamma_k = 0) = \tfrac{p}{2}\log\big(1 + n_k g_v\big) - \tfrac{1}{2}\chi^2_p(2n_k I_k) + \log\frac{\pi}{1-\pi}, \tag{16}$$

with mean $\frac{p}{2}\log(1 + n_k g_v) - \frac{p}{2} - n_k I_k$ and variance $\frac{p}{2} + 2n_k I_k$.

Three things follow that the fixed-alternative statement does not give. First, the Occam term $\frac{p}{2}\log(1+n_k g_v)$ is common to every source and does not involve $\boldsymbol{\delta}^{(k)}$ at all, which is the formal version of the observation in Section 5 that all ten sources contribute a large and nearly equal amount of evidence for inclusion and the compatibility signal is the spread on top. Second, the discriminating quantity between an informative and a noninformative source is the difference in noncentralities, $n_k(I_k^{\mathrm{non}} - I_k^{\mathrm{inf}})$, while the fluctuation is $\sqrt{p/2 + 2n_k I_k}$. Third, and this is the point of the exercise, a search over 2^K configurations requires the mean gap to dominate the maximum fluctuation, giving

$$n_k\big(I_k^{\mathrm{non}} - I_k^{\mathrm{inf}}\big) \;\gtrsim\; \sqrt{2\Big(\tfrac{p}{2} + 2n_k I_k\Big)K\log 2} \;+\; \log\frac{1-\pi}{\pi}. \tag{17}$$

Solving for n_k when $2n_k I_k \gg p$ gives $n_k \gtrsim 4K\log 2 \cdot I_k^{\mathrm{non}}/(I_k^{\mathrm{non}} - I_k^{\mathrm{inf}})^2$: linear in K and quadratic in the inverse rate gap. In the design of Section 5, $\|\boldsymbol{\delta}^{(k)}\|_2^2 = 4.32$ for noninformative sources and 0.18 for informative ones, so $I^{\mathrm{non}} = 2.16$, $I^{\mathrm{inf}} = 0.09$, and the requirement is $n_k \gtrsim 4(10)(0.693)(2.16)/(2.07)^2 \approx 14$, comfortably met. The contamination in that design is severe enough that separation is not the binding constraint; Section 5 shows the binding constraint is elsewhere.

Remark 8 (Where the orthogonal calculation fails)**.** Proposition 7 assumes $n_k \geq p$. When $n_k < p$ the source Gram matrix is singular, the excluded model is not identified, and (15) must be replaced by a chi-square on $r = \mathrm{rank}(X^{(k)}) = n_k$ degrees of freedom with noncentrality $n_k\|P_k\boldsymbol{\delta}^{(k)}\|^2/\sigma^2$, where P_k projects onto the row space. For a contrast in general position, $\mathbb{E}\|P_k\boldsymbol{\delta}^{(k)}\|^2 = (n_k/p)\|\boldsymbol{\delta}^{(k)}\|^2$, so the effective noncentrality falls to $n_k^2\|\boldsymbol{\delta}^{(k)}\|^2/(p\sigma^2)$, quadratic rather than linear in n_k. This degradation is real but, on the numbers above, still leaves the gap an order of magnitude above the union bound at $n_k = 150$. The failure we document at $n_k/p = 0.75$ in Table 1 is therefore not explained by the loss of power in (15), and we attribute it instead to the Occam term: when $n_k < p$ the estimation saving from pooling is of the same order as the misspecification cost for *every* source, so the posterior over $\boldsymbol{\gamma}$ is driven by $|\mathcal{A}_\gamma|$ rather than by its composition.

4.4 Multiplicity

The prior term in (17) is the multiplicity control. With π fixed at $1/2$ the term vanishes and the prior over $\boldsymbol{\gamma}$ is uniform on subsets, which supplies no adjustment as K grows: the expected number of included sources under the prior is $K/2$ regardless of K, and the prior odds for any single flip are one. Placing $\pi \sim \mathrm{Be}(a, b)$ and integrating gives

$$p(\boldsymbol{\gamma}) = \frac{B(a + m, b + K - m)}{B(a, b)}, \qquad m = \textstyle\sum_k \gamma_k, \tag{18}$$

whose log ratio for a single flip is $\log\{(a+m)/(b+K-m-1)\}$, reducing to $\log\{(m+1)/(K-m)\}$ at $a = b = 1$. This is the automatic multiplicity adjustment of Scott and Berger (2010): the penalty for

adding a source grows as the current model fills up, and at $m = K/2$ it is $\log\{(K/2+1)/(K/2)\} \approx 2/K$, small, while at $m = K - 1$ it is $\log K$. The adjustment therefore bites at the top of the model space, which is exactly where a procedure with a common Occam term across sources is liable to drift. Both the linear and logistic frameworks of Jamshidian (2026) fix the inclusion probability at $1/2$. Making it random costs nothing and is the correct default. We report in Section 5 that in the designs we examined the adjustment is inert, because the Bayes factors are far from the decision boundary; by (17) it will matter when K is large and the rate gap is small, which is the realistic genomic case.

4.5 Which specifications are affected

It is worth being precise about the scope of Proposition 4, because the three outcome settings in Jamshidian (2026) do not share a specification.

The Gaussian framework of Chapter 2 uses a single pooled contrast $\boldsymbol{\delta}$ for the whole informative block and a single pooled coefficient $\mathbf{w}^{(\bar{A})}$ for the whole excluded block. It is therefore exposed to both Proposition 4 and Proposition 6. This is the setting in which the posterior over configurations is stated most cleanly, in which the Bayes factor theory is proved, and in which the Bayesian model averaging interpretation is exact. It is also the setting with the specification problem.

The logistic framework of Chapter 3 replaces the pooled contrast by source-specific contrasts $\boldsymbol{\delta}^{(k)}$, each with its own global scale τ_k, and classifies sources by the shrinkage regime of τ_k. Structurally this is the source-specific nuisance specification, and it is immune to Proposition 4. The Cox model of Chapter 4 does the same, with the additional refinement of study-specific baseline hazards coupled through a matrix-normal prior. So the later chapters, adopted for reasons of computational tractability rather than of identification, happen to get the structure right.

They pay for it in a different currency. Once compatibility is defined by the shrinkage regime of τ_k rather than by a marginal likelihood comparison, the estimand changes. The posterior $\Pr(\gamma_k = 1 \mid \mathcal{D})$ then answers the question "is the contrast for study k shrunk more strongly than the target's own coefficients", which coincides with predictive compatibility only under the working prior and only through the separation constant a. The dissertation is explicit that a is estimated by empirical Bayes and that a lower bound on the global shrinkage parameter is imposed to repair undercoverage. Both are reasonable engineering choices, but together they mean that the reported inclusion probabilities in Chapters 3 and 4 are not comparable to those in Chapter 2 and are not estimating the same functional. Our recommendation is to keep the Chapter 2 estimand and fix the nuisance block, rather than to keep the nuisance block and change the estimand; Section 7 shows the first route is open for binary outcomes.

4.6 A diagnostic

Figure 2 and Proposition 7 together suggest a reporting convention. The log Bayes factor for including source k decomposes into an Occam term common to all sources and a compatibility term that is not:

$$\log \mathrm{BF}(\gamma_k = 1 : \gamma_k = 0) \;=\; \underbrace{\tfrac{p}{2}\log(1 + n_k g_v)}_{\text{pooling efficiency}} \;-\; \tfrac{1}{2}\underbrace{\chi^2_p\big(2n_k I_k\big)}_{\text{incompatibility}}. \tag{19}$$

The first term is source independent and large: at $p = 200$, $n_k = 300$, $g_v = 1$ it is 571 nats, against a compatibility term with mean $p/2 + n_k I_k = 127$ nats for an informative source. Figure 2 shows the empirical consequence in the case $|\mathcal{A}_\gamma| = 1$, where Proposition 7 does not apply exactly because the anchor is not estimated with negligible error: the realized log Bayes factors cluster near 140 nats for every source, informative or not, and the discriminating spread is a few nats on top. Reporting only the

		$\mathbb{E}[\Pr(\gamma_k = 1 \mid \mathcal{D})]$		
n_k/p	nuisance block	$k \in \mathcal{A}$	$k \notin \mathcal{A}$	mean $\lvert\cdot\rvert_1$ error
0.75	pooled (BLAST)	1.000	0.598	0.418 (0.027)
0.75	source-specific	1.000	0.576	0.403 (0.021)
1.50	pooled (BLAST)	0.995	0.666	0.468 (0.019)
1.50	source-specific	1.000	0.143	0.100 ($< 10^{-4}$)
3.00	pooled (BLAST)	0.996	0.665	0.467 (0.019)
3.00	source-specific	1.000	0.000	0.000

Table 1: Exact posterior inclusion probabilities from enumeration of all 2^{10} configurations, averaged over six replicates, standard errors in parentheses. The $\lvert\cdot\rvert_1$ error is the mean absolute deviation of the inclusion probability from the truth. The pooled nuisance block does not improve with sample size.

posterior inclusion probability hides this, since it reports the sign of a difference between two quantities of very different magnitude.

Proposition 7 supplies the missing summary directly. Since $\mathbb{E}[\chi^2_p(\lambda_k)] = p+\lambda_k$ and $\mathrm{Var} = 2p+4\lambda_k$, the statistic

$$\widehat{I}_k = \frac{1}{2n_k}\Big\{-2\Delta\ell_k - p\Big\}, \qquad \mathrm{sd}\big(\widehat{I}_k\big) = \frac{\sqrt{2p + 8n_k I_k}}{2n_k}, \tag{20}$$

is an unbiased estimator of the Sanov rate I_k of (8) with a standard error available in closed form, and $\Delta\ell_k$ is a by-product of the sampler at no extra cost. Two features make it the right thing to report. It is on the scale of nats per observation, so it is comparable across sources of different size, unlike $\log \mathrm{BF}$ itself. And its standard error is $O(\sqrt{p}/n_k)$ in the small-rate regime, which makes visible the fact that a rate of order p/n_k^2 is not estimable at all, independently of how confident the inclusion probability looks.

We suggest reporting $\widehat{I}_k \pm 2\,\mathrm{sd}(\widehat{I}_k)$ alongside $\Pr(\gamma_k = 1 \mid \mathcal{D})$, together with the two terms of (19) separately. The comparison to make is between $n_k\widehat{I}_k$ and $p/2$: by (15) the compatibility term has mean $p/2 + n_k I_k$, so a source whose rate satisfies $n_k I_k \lesssim p/2$ contributes no more than the null fluctuation, and its inclusion probability is determined by the Occam term rather than by its data. In the TCGA analysis of the source material, $p = 303$ genes sets that floor at about 150 nats, against an Occam term of $\frac{p}{2}\log(1 + n_k g_v) \approx 860$ nats at $n_k = 300$.

5 Exact enumeration

The claims above concern the posterior $p(\boldsymbol{\gamma} \mid \mathcal{D})$, not an algorithm. To separate the statistical question from the mixing of Metropolis-within-Gibbs, we enumerate all 2^K configurations exactly in a conjugate version of the model in which σ^2 is known and the scale mixtures are replaced by ridge priors, $\mathbf{w} \sim \mathcal{N}(0, \sigma^2 g_w I)$, $\boldsymbol{\delta} \sim \mathcal{N}(0, \sigma^2 g_\delta I)$, and the nuisance block likewise. The marginal likelihood is then available in closed form for each $\boldsymbol{\gamma}$, at the cost of one $2p \times 2p$ and one $p \times p$ Cholesky factorization.

The design follows the reference simulations: $K = 10$, $p = 200$, $n_0 = 150$, three informative sources with $\mathbf{w}^{(k)} = \boldsymbol{\beta} - 0.3\,\mathbb{I}(j \in H_k)$, $|H_k| = 2$, and seven noninformative sources with deviation 0.6 on $|H_k| = 12$ coordinates, with $\boldsymbol{\beta}$ having six signals of size 0.5. We vary n_k and the nuisance specification, using $g_w = 1$, $g_\delta = 0.1$, $\pi = 1/2$, and six replicates.

Table 1 and Figure 1 confirm Propositions 4 and 5. Under the pooled nuisance block the expected inclusion probability of a noninformative source is 0.60, 0.67, 0.66 at $n_k/p = 0.75, 1.5, 3.0$: it does not decay, and at $n_k/p = 3$ the procedure is no better than at $n_k/p = 0.75$. Under the source-specific block

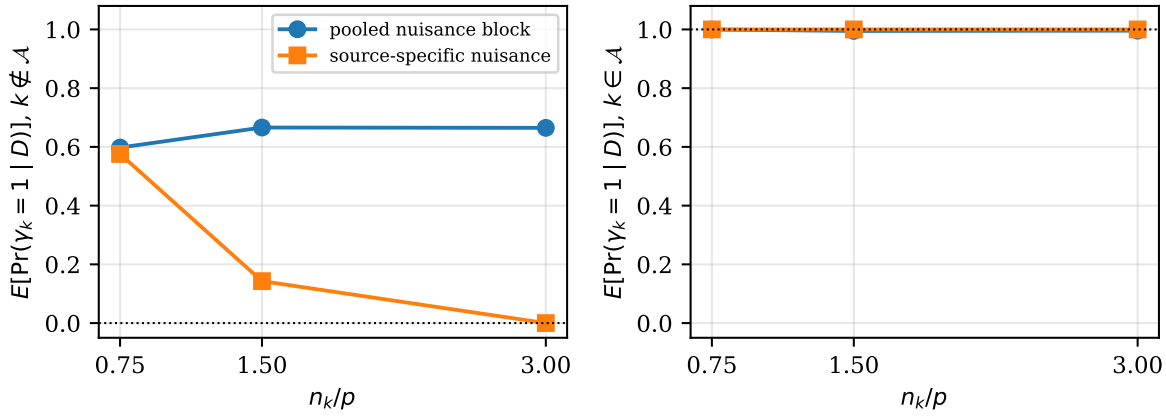

Figure 1: Exact posterior inclusion probabilities against n_k/p for noninformative sources (left) and informative sources (right). Informative sources are recovered under both specifications. Noninformative sources are excluded consistently only under the source-specific nuisance block, and only once n_k exceeds p.

the same quantity is $0.58, 0.14, 0.00$. Informative sources are recovered with probability one throughout, so the failure is entirely one of false inclusion.

Two further observations. First, at $n_k/p = 0.75$ the two specifications are indistinguishable and both fail, exactly as Proposition 5 predicts: below $n_k = p$ the exclusion alternative is itself unidentified and pooling always wins. The values we obtain, close to 1.0 for informative and close to 0.6 for noninformative sources, are a close match to the 0.7 against 0.45 pattern reported at $n_k = 150$, $p = 200$ in Jamshidian (2026, Fig. 2.2). Since our enumeration is exact, this establishes that the weak separation observed there is a property of the posterior and not an artifact of the sampler or of the tempering schedule. Second, at $n_k/p = 1.5$ the source-specific specification retains exactly one false inclusion in each of the six replicates, giving the $0.143 = 1/7$ in Table 1. We do not have a complete explanation for the persistence of exactly one; it is consistent with a regime in which the estimation saving in Proposition 5 covers one additional source but not two.

Figure 2 isolates the identification result. At $n_k = 300$ and with a diffuse contrast prior $g_\delta = 1$, the evidence gap when the informative pool would contain a single source is 0.40 nats. Tightening the contrast prior to $g_\delta = 0.01$ raises it to 34.2 nats. Once the pool contains two sources the gap is 152.9 nats at $g_\delta = 1$ and 171.0 at $g_\delta = 0.01$; with three sources, 302.0 and 315.4. The pattern is exactly Proposition 6: at $|\mathcal{A}_\gamma| = 1$ the discrimination is supplied entirely by the contrast prior and disappears as that prior is made diffuse, whereas at $|\mathcal{A}_\gamma| \geq 2$ it is supplied by mutual coherence and is enormous. The right panel makes the scale explicit. Every source, informative or not, contributes about 140 nats of evidence for inclusion, because $n_k = 300$ observations always help estimate a 200-dimensional coefficient vector. The compatibility signal is the few nats of spread on top of that.

We also tested the multiplicity adjustment of Section 4. Replacing $\pi = 1/2$ by $\pi \sim \mathrm{Be}(1, 1)$ changed the expected number of false inclusions from 0.99997 to 0.99992 at $n_k/p = 1.5$. In this design the adjustment is inert, because the flips are decided by hundreds of nats and the prior contributes about one. We report this as a negative result. It does not contradict (17); it says that the designs used to benchmark these methods are not in the regime where multiplicity binds, which is itself informative about the benchmarks.

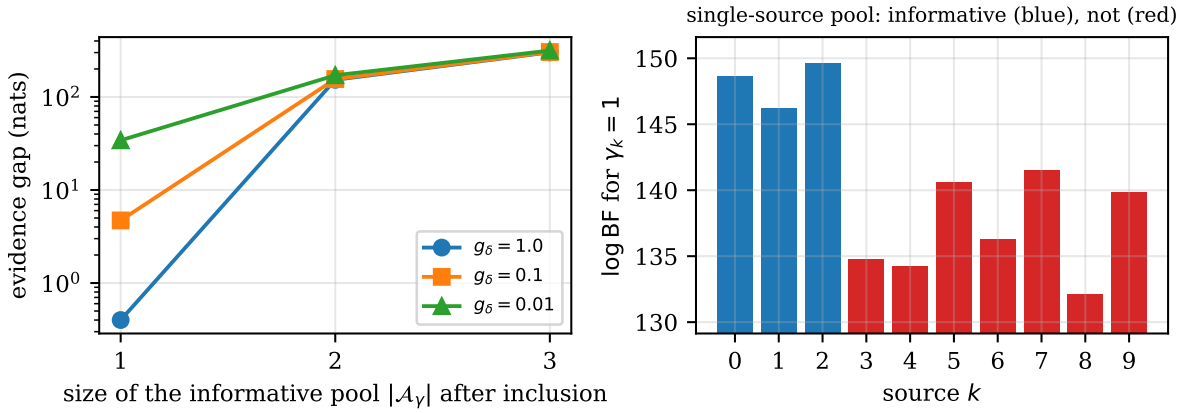

Figure 2: Left: the evidence gap, defined as the smallest log Bayes factor for including a truly informative source minus the largest for including a noninformative one, against the size of the informative pool after inclusion, at $n_k = 300$. Right: log Bayes factors for each candidate when the pool would contain that source alone. The overall magnitude is around 140 nats and the discriminating component is a few nats.

		pooled nuisance		source-specific nuisance	
σ^2	n_k/p	$k \in \mathcal{A}$	$k \notin \mathcal{A}$	$k \in \mathcal{A}$	$k \notin \mathcal{A}$
known	1.5	0.995	0.666	1.000	0.143
known	3.0	0.996	0.665	1.000	0.000
estimated	1.5	0.000	0.000	1.000	1.000
estimated	3.0	0.043	0.137	1.000	0.000

Table 2: Expected posterior inclusion probabilities with the block variances known and integrated out. Estimating the variances does not repair the pooled specification; it replaces one pathology with a worse one, and it costs the source-specific specification a factor of two in the sample size needed for recovery.

5.1 Sensitivity to the variance specification

The enumeration above fixes $\sigma^2 = 1$, which is not what the framework does: each block carries its own residual variance with an $\mathrm{IG}(\omega/2, \omega/2)$ prior. A pooled nuisance block that fits its sources badly therefore has a second way out. Instead of relocating a source into the informative pool, it can inflate $\sigma^2_{(\bar{\mathcal{A}})}$ and call the misfit noise. Whether that repairs Proposition 4 is a fair question and it is answerable in closed form, since with a normal-inverse-gamma prior the variance integrates out analytically and the enumeration remains exact.

We repeated the experiment with each block's variance integrated out, one variance for the informative block and either one shared or $|\bar{\mathcal{A}}_\gamma|$ separate variances for the excluded sources, taking $\omega = 1$.

Table 2 says that the extra flexibility does not help and in one case inverts the answer. Under the pooled block at $n_k/p = 1.5$ the posterior excludes everything, with inclusion probability numerically zero for informative and noninformative sources alike: the nuisance block absorbs all seven noninformative sources by inflating its variance, and having done so it absorbs the three informative ones too, so no transfer occurs at all. At $n_k/p = 3$ the ordering inverts outright, with noninformative sources receiving inclusion probability 0.137 against 0.043 for informative ones. A procedure that prefers the sources it

should reject is worse than one that cannot tell them apart, and we did not anticipate this.

The source-specific block is not unaffected either. With σ^2 known it separates at $n_k/p = 1.5$; with σ^2 estimated it includes everything at that ratio and requires $n_k/p = 3$ to recover $\mathcal{A}$ exactly. Estimating p-dimensional coefficient vectors and a variance from n_k observations leaves the excluded model too flexible at moderate n_k, and inclusion wins by default. The direction of the effect is the same in both columns: freeing the variances makes selection harder, not easier.

Two caveats, both material. Our variance hierarchy is a simplification of the framework's. We assign one variance to the whole informative block, target and included sources together, whereas the original carries separate $\sigma^2_{(\mathcal{A})}$ and $\sigma^2_{(0)}$ and scales the priors on $\mathbf{w}$ and $\boldsymbol{\delta}$ by their respective block variances, a coupling that does not admit a closed-form marginal. So Table 2 establishes that the selection posterior is acutely sensitive to how the variances are specified, which we do not think has been noted, but it does not establish that the published hierarchy inverts. Second, three replicates at $n_k/p = 3$ is thin for a claim about an inversion, and we would want more before resting anything on it.

What the table does settle is the direction of the earlier open question. Giving the model more freedom to explain misfit does not rescue a misspecified nuisance block. It supplies another channel through which the block can avoid identifying $\mathcal{A}$, and the pooled specification uses it.

Finally, we checked for a look-elsewhere pathology under an exchangeable null in which all ten sources are informative with identically distributed contrasts, so that no true partition exists. At $n_k/p = 1.5$ the posterior concentrated on the all-inclusive configuration with zero expected spurious exclusions and posterior entropy numerically indistinguishable from zero. The procedure does not manufacture a partition when none is present. The failure mode is one-sided: false inclusion, not false discovery of structure.

6 Does any of this affect target inference?

Sections 4 and 5 establish that the posterior over configurations can be badly wrong and that the error does not diminish with sample size. A referee is entitled to ask whether the target inference notices. The answer turns out to depend entirely on the target sample size, and the two regimes point in opposite directions. At the sample sizes used in the published benchmarks the target barely notices, which we show first because it is the more surprising half and because it indicts the benchmarks. At the sample sizes for which transfer learning is actually reached for, the target notices a great deal, and the specification defect of Proposition 4 costs a factor of three. Section 6.3 gives that half.

6.1 Point estimation

Table 3 reports the sum of squared error for the posterior mean of $\boldsymbol{\beta}$ in the design of Section 5 at $n_k = 300$, that is at $n_k/p = 1.5$, where the two nuisance specifications differ maximally in their inclusion probabilities (0.666 against 0.143 for noninformative sources). Five estimators: the target-only analysis, the oracle that conditions on the true $\mathcal{A}$, the naive analysis that includes all ten sources, and the two Bayesian model averages.

Two things stand out. Transfer is worth a great deal: the oracle more than halves the target-only error. Selection is worth almost nothing: including all ten sources, seven of which are noninformative with contrasts three times the size of the informative ones, costs 7% relative to the oracle. The two model averages are indistinguishable from each other to three significant figures, at 1.666 and 1.665, despite assigning noninformative sources posterior inclusion probabilities that differ by 0.52.

estimator	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2^2$	relative to oracle
target only	3.353 (0.174)	2.13
oracle, $\mathcal{A}$ known	1.572 (0.042)	1.00
all sources included	1.686 (0.038)	1.07
model average, pooled nuisance	1.666 (0.044)	1.06
model average, source-specific	1.665 (0.048)	1.06

Table 3: Target estimation error at $n_k/p = 1.5$, six replicates, standard errors in parentheses. Transfer halves the error relative to a target-only analysis. Which sources are transferred from is nearly immaterial.

The mechanism is the one identified in Proposition 6, read in the other direction. The anchor $\mathbf{w}$ plus a free contrast $\boldsymbol{\delta}$ can absorb a great deal of source contamination before the target estimate degrades, because the contrast has p free coordinates and is estimated from the target data. What protects the target from negative transfer is the contrast parameterization itself, not the selection step layered on top of it. The selection step is a refinement of something that is already robust.

6.2 Uncertainty quantification

The stronger claim in the source material is not about point estimation but about intervals: that averaging over $\boldsymbol{\gamma}$ propagates selection uncertainty into the target credible intervals in a way that conditioning on $\hat{\mathcal{A}}$ does not. We computed the exact model-averaged posterior variance by the law of total variance,

$$\mathrm{Var}(\beta_j \mid \mathcal{D}) = \sum_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma} \mid \mathcal{D}) \, \mathrm{Var}(\beta_j \mid \mathcal{D}, \boldsymbol{\gamma}) + \sum_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma} \mid \mathcal{D}) \left\{ \mathbb{E}(\beta_j \mid \mathcal{D}, \boldsymbol{\gamma}) - \mathbb{E}(\beta_j \mid \mathcal{D}) \right\}^2, \tag{21}$$

where the second term is exactly the selection-uncertainty contribution. Averaged over replicates at $n_k/p = 1.5$, the mean 95% interval width for signal coordinates is 0.7365 under the pooled nuisance specification and 0.7378 under the source-specific one, a difference of 0.17%; for non-signal coordinates, 0.7294 against 0.7307. Empirical coverage is 1.00 in both cases, reflecting the conservatism of the ridge prior used in the enumeration.

So the second term in the variance decomposition, the one that carries the entire conceptual justification for averaging over $\boldsymbol{\gamma}$, is numerically negligible here. The configurations being averaged over have target posteriors that are nearly identical, so averaging over them adds nearly nothing. This is not an argument against averaging, which is free and correct. It is an argument against expecting much from it in this design, and against using interval width as a diagnostic for whether the selection step is working.

6.3 A design in which selection is load-bearing

The designs above are inherited from the literature we are examining, and they share a feature that makes the negative result almost automatic: a target sample of $n_0 = 150$ with only six signals. The contrast $\boldsymbol{\delta}$ is estimated from the target alone. If n_0 is comfortable relative to the number of signals, $\boldsymbol{\delta}$ is estimated well, and it can absorb whatever the anchor gets wrong. Transfer learning, however, exists for the case where n_0 is not comfortable. We therefore repeated the experiment varying n_0, with noninformative sources deviating by 1.0 on twelve coordinates and a tight contrast prior $g_\delta = 0.05$, the regime the truncation $\tau_{(0)} < \tau_{(\mathcal{A})}$ is designed to enforce.

	$n_0 = 50$		$n_0 = 100$		$n_0 = 300$	
estimator	SSE	ratio	SSE	ratio	SSE	ratio
target only	1.535	2.61	1.797	2.32	1.884	1.79
oracle, $\mathcal{A}$ known	0.587	1.00	0.776	1.00	1.053	1.00
all sources included	1.709	2.91	1.479	1.91	1.098	1.04
model average, pooled nuisance	1.780	3.03	1.609	2.07	1.077	1.02
model average, source-specific	0.587	1.00	0.776	1.00	1.053	1.00

Table 4: Target estimation error against target sample size, with severe contamination ($\|\boldsymbol{\delta}^{(k)}\|_\infty = 1.0$ on twelve coordinates for seven of ten sources) and a tight contrast prior. Six replicates, $n_k = 300$, $p = 200$. The ratio is to the oracle. At $n_0 = 50$ including all sources is worse than ignoring them, and the pooled nuisance specification is worse still.

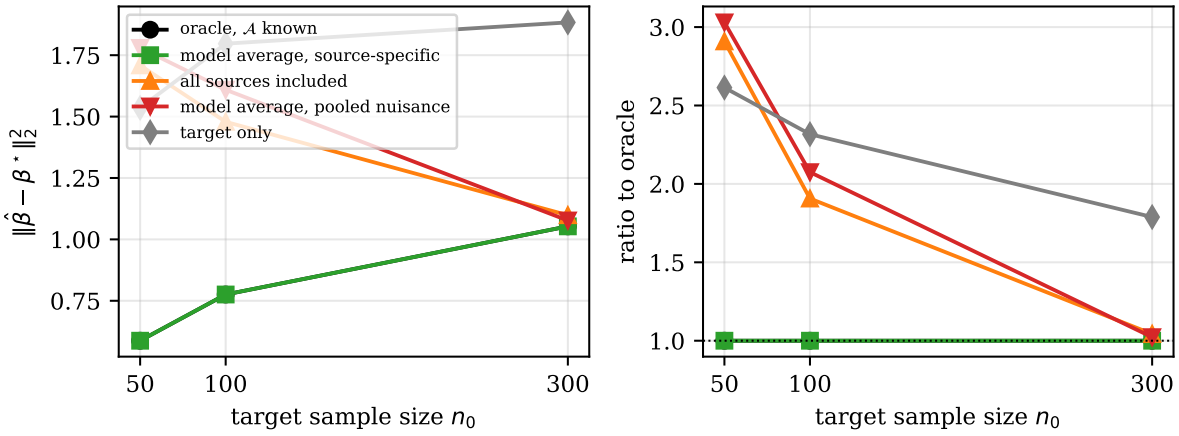

Figure 3: Target estimation error (left) and ratio to the oracle (right) against target sample size, under severe contamination. The cost of getting selection wrong is a factor of three at $n_0 = 50$ and two percent at $n_0 = 300$. The source-specific model average attains the oracle exactly at every n_0, because it recovers $\mathcal{A}$ with posterior probability one.

6.4 An absorption factor

The pattern in Table 4 has an explanation that can be made exact, and the exact version differs from its isotropic caricature by an order of magnitude, which is itself the point.

Proposition 9 (Contamination absorption)**.** *Consider the conjugate model with ridge scales g_w, g_δ and informative pool $\mathcal{A}_\gamma$. Write $S_0 = X^{(0)\top}X^{(0)}$, $S_\mathcal{A} = \sum_{k\in\mathcal{A}_\gamma} X^{(k)\top}X^{(k)}$, $C = S_0 + g_\delta^{-1}I$, and let*

$$M \;=\; \sum_{k\in\mathcal{A}_\gamma} X^{(k)\top}X^{(k)}\boldsymbol{\delta}^{(k)} \tag{22}$$

be the design-weighted *aggregate contrast. Then the posterior mean of the target satisfies, exactly and for any designs,*

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] - \boldsymbol{\beta} \;=\; -\frac{1}{g_\delta}\,C^{-1}\big(S_\mathcal{A} + S_0 + g_w^{-1}I - S_0C^{-1}S_0\big)^{-1}\big(M + g_w^{-1}\boldsymbol{\beta}\big). \tag{23}$$

If in addition $S_0 = n_0 I$ and $g_w \to \infty$, this collapses to

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] - \boldsymbol{\beta} \;=\; -\frac{1}{\zeta}\Big(S_{\mathcal{A}} + \frac{n_0}{\zeta} I\Big)^{-1} M, \qquad \zeta = 1 + n_0 g_\delta. \tag{24}$$

Corollary 10 (Isotropic case)**.** *If furthermore $X^{(k)\top} X^{(k)} = n_k I$ for all k, then $M = n_k \Delta$ with $\Delta = \sum_{k \in \mathcal{A}_\gamma} \boldsymbol{\delta}^{(k)}$, and for $|\mathcal{A}_\gamma| n_k \gg n_0$*

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] - \boldsymbol{\beta} \;\approx\; -\frac{\Delta}{|\mathcal{A}_\gamma|\,(1 + n_0 g_\delta)}. \tag{25}$$

Equation (25) is the interpretable form. The contrast soaks up all but a fraction $\{|\mathcal{A}_\gamma|(1 + n_0 g_\delta)\}^{-1}$ of whatever the anchor gets wrong, and its capacity to do so grows linearly in $n_0 g_\delta$. Selection is therefore inferentially load-bearing when

$$\frac{\|\Delta\|_2^2}{|\mathcal{A}_\gamma|^2 (1 + n_0 g_\delta)^2} \;\gtrsim\; \big\|\hat{\boldsymbol{\beta}}_{\text{oracle}} - \boldsymbol{\beta}\big\|_2^2, \tag{26}$$

which is the threshold we lacked. Note what controls it: not n_0 alone but the product $n_0 g_\delta$. A tighter contrast prior, imposed for identification as in Section 2.2, simultaneously makes the target *more* exposed to bad sources. Identification and robustness are governed by the same number and pull against each other. We have not seen this coupling noted.

Table 5 checks both versions against the simulation, and the gap between them is instructive. The isotropic corollary reproduces the observed twenty-five-fold decline in excess error from $n_0 = 50$ to $n_0 = 300$ almost exactly, with the intermediate point out by a factor of two, so it is a serviceable scaling law. Its absolute level is wrong by a factor of fourteen: with $\|\Delta\|_2^2 = 121$ in this design, (26) predicts an excess of 0.098 against the observed 1.121.

Evaluating (23) at the realized design matrices instead gives a squared bias of 1.379 under naive pooling and 0.071 under the oracle, with total posterior variances 7.839 and 8.000, hence a predicted excess of $(1.379 - 0.071) + (7.839 - 8.000) = 1.147$ against the observed 1.121. The formula is therefore correct and the caricature is what fails. Two features of the isotropic approximation are responsible, and the second is the larger. Replacing M by $n_k \Delta$ discards the design fluctuation $\sum_k (X^{(k)\top} X^{(k)} - n_k I)\boldsymbol{\delta}^{(k)}$, which is not mean-zero in norm and contributes at order $\sqrt{p\, n_k}\, \|\boldsymbol{\delta}^{(k)}\|$ per source. Replacing $S_{\mathcal{A}}$ by $|\mathcal{A}_\gamma| n_k I$ discards the Marchenko-Pastur spread of the pooled Gram matrix, which at $p/(|\mathcal{A}_\gamma| n_k) = 0.067$

	$n_0 = 50$	$n_0 = 100$	$n_0 = 300$
isotropic factor (25)	0.0286	0.0167	0.0063
predicted excess, isotropic (26)	0.098	0.033	0.0047
predicted excess, exact (23) plus variance	1.147	—	—
observed excess, Table 4	1.121	0.703	0.045
predicted scaling, relative to $n_0 = 50$	1.00	0.34	0.048
observed scaling, relative to $n_0 = 50$	1.00	0.63	0.040

Table 5: The absorption factor against the observed excess error of naive pooling over the oracle. The isotropic corollary gives the scaling in n_0 but understates the level by an order of magnitude; the exact formula (23) evaluated at the realized designs predicts the excess to within 2%.

runs over $[1652, 4744]$, so contamination aligned with the small eigenvalues is amplified. Together these account for the factor of fourteen.

The variance figures are worth noting separately. At 7.839 against 8.000, the estimation saving from including seven extra sources is 2%, while the bias cost is a factor of nineteen. In this regime pooling buys essentially nothing and risks a great deal, which is the quantitative statement of why selection matters at small n_0.

Table 4 and Figure 3 answer the question left open above. At $n_0 = 50$ the transfer is genuinely dangerous: including all ten sources gives SSE 1.709 against 1.535 for ignoring the sources entirely, so naive pooling is worse than no transfer at all. This is negative transfer in the strict sense, and it is exactly what the selection machinery is for. The pooled nuisance specification does not deliver: at 1.780 it is worse than naive pooling, and three times the oracle. The source-specific specification recovers the informative set with posterior probability one, and therefore attains the oracle error exactly, 0.587.

As n_0 grows the gap closes. At $n_0 = 300$ the pooled specification costs 2% and naive pooling costs 4%, which is the regime of Table 3 and of the published benchmarks. The posterior inclusion probabilities are essentially unchanged across the three columns, at 0.59, 0.62, 0.60 for noninformative sources under the pooled specification and numerically zero under the source-specific one. The selection posterior is equally wrong throughout; what changes is whether being wrong costs anything.

We therefore withdraw the strongest form of the negative result. Source selection is inferentially load-bearing, and the specification defect of Proposition 4 carries a factor of three, in precisely the regime where transfer learning is used rather than benchmarked. The reason it has gone unnoticed is that the standard simulation designs use a target sample large enough for the contrast to absorb the damage.

6.5 What follows

Taking Sections 6.1 to 6.3 together, the picture is this. The cost of a wrong selection posterior is governed by how well the target can estimate its own contrast, which is governed by n_0 relative to the number of signals. When n_0 is generous the contrast absorbs the contamination and selection is close to free, which is the regime of every benchmark we are aware of. When n_0 is small the contrast cannot absorb it, naive pooling produces negative transfer, and the selection step is doing the work it was designed to do. The specification of the nuisance block then determines whether it succeeds or fails, and in our design it is the difference between attaining the oracle and tripling its error.

Three implications for practice. First, benchmark at the n_0 at which the method will be used. A simulation with $n_0 = 150$ and six signals cannot distinguish a working selection step from a broken one, and a method can fail selection completely while scoring well on estimation error, prediction error, interval width and coverage simultaneously. Second, report the diagnostic of Section 4.6 rather than relying on those four quantities. Third, when posterior inclusion probabilities are being read substantively, as claims that one cancer shares molecular architecture with another, they should be treated as the fragile output they are: our Propositions 4 and 6 say they can be wrong even when every reported performance metric looks healthy.

7 Collapsed marginals under Pólya-Gamma augmentation

For binary outcomes the marginal likelihood needed by the selection step is not available in closed form, and Jamshidian (2026) responds by abandoning marginal likelihood comparison entirely, replacing it with a two-component mixture on the study-level global shrinkage parameter: with $\xi_k = \tau_k^{-2}$,

$$\xi_k \mid \gamma_k, \tau_0 \sim \gamma_k \,\mathrm{Ga}\big(1, 1/(a\xi_0)\big) + (1-\gamma_k)\,\mathrm{Ga}\big(1, a/\xi_0\big), \qquad a > 1, \tag{27}$$

so that $\mathbb{E}[\xi_k \mid \gamma_k = 0] < \xi_0 < \mathbb{E}[\xi_k \mid \gamma_k = 1]$ and, marginally, $\delta_j^{(k)}$ is a two-component mixture of t_2 scales. This is elegant and conjugate. It is also a different estimand: compatibility is now defined by the shrinkage regime of the contrast rather than by predictive agreement, and the two coincide only under the working prior.

The step is avoidable. Under Pólya-Gamma augmentation (Polson et al., 2013), with $\omega_i^{(k)} \sim \mathrm{PG}(m_i^{(k)}, \psi_i^{(k)})$, $\kappa = y - m/2$ and working response $z = \Omega^{-1}\kappa$, the model is conditionally Gaussian, so the marginal likelihood *given* ω is exactly the Gaussian marginal with heteroscedastic weights,

$$m(y \mid \boldsymbol{\gamma}, \omega) = |\Omega|^{1/2}(2\pi)^{-n/2}\frac{\exp\{-\frac{1}{2}(z^\top \Omega z - \mu^\top \Lambda \mu)\}}{|G|^{1/2}|\Lambda|^{1/2}}, \qquad \Lambda = Z_\gamma^\top \Omega Z_\gamma + G^{-1}, \tag{28}$$

with $\mu = \Lambda^{-1} Z_\gamma^\top \Omega z$ and G the prior scale matrix. Two schemes follow. The first is a collapsed Gibbs step: update $\boldsymbol{\gamma}$ from $p(\boldsymbol{\gamma} \mid \omega, \nu, \mathcal{D}) \propto m(y \mid \boldsymbol{\gamma}, \omega)\, p(\boldsymbol{\gamma})$, with $\mathbf{w}$ and $\boldsymbol{\delta}$ integrated out analytically. This is exact for the augmented target and preserves the Chapter 2 estimand. The second is pseudo-marginal: form an unbiased estimate of $m(y \mid \boldsymbol{\gamma})$ by importance sampling over ω with the Pólya-Gamma prior as proposal, and use it in a Metropolis step, which is exact for the collapsed target by the usual argument. Neither requires the shrinkage-regime reformulation, and both keep the Bayesian model averaging in (2) over the same object for Gaussian and binary outcomes. Everything in Sections 4 and 5 then applies verbatim, including the warning about the nuisance specification.

8 Admissible mixing measures and the Thorin class

The scale mixture representation $w_j \mid \nu_j \sim \mathcal{N}(0, \sigma^2\nu_j)$, $\nu_j \sim f$, leaves f free. For the selection step, the relevant question is which f leave the marginal likelihood tractable. Write the marginal density of w_j as

$$p(w_j) = \int_0^\infty \mathcal{N}(w_j \mid 0, \sigma^2\nu)\, f(\nu)\, d\nu, \tag{29}$$

a Gaussian scale mixture, so p is determined by the Laplace transform of f evaluated at $w_j^2/(2\sigma^2)$. If f belongs to the Thorin class, that is if ν is a generalized gamma convolution with Thorin measure U, then

$$\mathbb{E}\big[e^{-s\nu}\big] = \exp\Big\{-as - \int_0^\infty \log\big(1 + s/u\big)\, U(du)\Big\}, \tag{30}$$

and the marginal is available in closed form up to the one-dimensional integral defining the Thorin measure. The horseshoe, the Bayesian lasso, the normal-gamma and the normal-inverse-Gaussian all sit in this class. Two payoffs. First, the log marginal likelihood of a configuration, which is the object driving selection, inherits a representation as a Thorin integral, so the pairwise comparisons of Section 4 can be carried out for the whole family at once rather than prior by prior. Second, self-decomposability of the mixing law is exactly the condition under which the study-level and coefficient-level shrinkage compose, that is under which placing a mixture on τ_k and a mixture on λ_{kj} yields a marginal still in the class. This is what makes the two-level construction of Section 2 coherent rather than merely convenient, and we regard the characterization of the admissible pairs as the natural next technical step.

9 Generative computation

The computational obstacle in this literature is stated plainly in the source material. The Gaussian sampler needs burn-in tempering to mix over $\boldsymbol{\gamma}$; the survival model abandons global-local shrinkage for a

Laplace prior purely for stability under Hamiltonian Monte Carlo, and its simulation study is capped at $K = 5$, $p = 100$ and twenty replicates.

This is the regime for generative Bayesian computation. The model is easy to simulate from and hard to integrate. Draw $\theta^{(m)} = (\boldsymbol{\beta}, \boldsymbol{\delta}^{(1:K)}, \boldsymbol{\gamma}, \nu) \sim \pi(\theta)$ and $\mathcal{D}^{(m)} \sim p(\mathcal{D} \mid \theta^{(m)})$ for $m = 1, \ldots, M$, reduce each synthetic dataset to a summary $S(\mathcal{D}^{(m)})$, and learn the conditional quantile map $\tau \mapsto Q_{\boldsymbol{\beta}|S}(\tau)$ by deep quantile regression. Posterior draws are then obtained by evaluating the learned map at uniform inputs, and $\boldsymbol{\gamma}$ is never instantiated: it is integrated out in the simulation rather than sampled. The configuration space that forces tempering in MCMC costs nothing here. The obvious summary is the vector of per-source sufficient statistics $\{(X^{(k)\top}X^{(k)}, X^{(k)\top}y^{(k)})\}$, which is exactly sufficient in the Gaussian case, and a set of partial-likelihood scores in the Cox case. We regard a generative implementation of the survival model, restoring horseshoe shrinkage and lifting the dimension caps, as the most useful piece of applied work suggested by this analysis.

10 Discussion

The Bayesian formulation of multi-source transfer learning is right in its main commitment: uncertainty about which sources are informative should propagate into target inference, and (2) is how that is done. Our results concern the object being averaged over.

10.1 When selection matters

The two halves of Section 6 pull in opposite directions and it is worth saying plainly what we take from them.

The contrast parameterization is a genuine robustness device, and this deserves more credit than it has received. A free p-dimensional deviation estimated from the target absorbs a great deal of source contamination, which is why naive pooling performs respectably in the published simulations and why the elaborate selection machinery buys only a few percent there. Anyone reporting that a selection method beats naive pooling by a small margin should check whether the margin is small because selection is unnecessary in their design.

But the absorption has a budget, and the budget is the target sample size relative to the signal count. Section 6.3 exhausts it by taking $n_0 = 50$, at which point naive pooling is worse than discarding the sources and the ranking of methods inverts. We do not think $n_0 = 50$ is an artificial choice. It is the situation in a rare cancer cohort, in a single-site trial, or in any of the applications that motivate borrowing in the first place. The benchmarks are run at $n_0 = 150$ because that is where the competitor methods are stable, not because that is where the method will be used.

The uncomfortable consequence is that the four standard evaluation metrics are jointly blind to selection failure at the sample sizes at which they are computed. A method can assign noninformative sources posterior inclusion probability 0.6, be provably inconsistent for $\mathcal{A}$, and still return estimation error, prediction error, interval width and coverage indistinguishable from a correct procedure. That is what the pooled nuisance specification does in Table 3, and it is why the defect has not been caught.

10.2 The survival case: a confounded rate function

We have said little about time-to-event outcomes and should say why, because the mathematics is worse there in a way that is easy to miss.

The Cox formulation in the source material carries a second nuisance layer, the study-specific baseline hazards, coupled through a matrix-normal prior on spline coefficients. Write the source and target

intensities as $\lambda_k(t \mid x) = h_0^{(k)}(t)\exp\{x^\top(\boldsymbol{\beta} - \boldsymbol{\delta}^{(k)})\}$ and $\lambda_0(t \mid x) = h_0^{(0)}(t)\exp(x^\top\boldsymbol{\beta})$, and set

$$\psi_k(t) = \log h_0^{(k)}(t) - \log h_0^{(0)}(t), \qquad \eta_k(x) = x^\top\boldsymbol{\delta}^{(k)}, \tag{31}$$

for the baseline and coefficient discrepancies. The relative entropy between the two counting-process laws over a follow-up window $[0, \bar{t}]$ with at-risk indicator $Y(t)$ is, by the Andersen-Gill representation,

$$I_k^{\mathrm{cox}} = \mathbb{E}\int_0^{\bar{t}} \Big[\big\{\psi_k(t) - \eta_k(x)\big\}\lambda_k(t \mid x) - \lambda_k(t \mid x) + \lambda_0(t \mid x)\Big] Y(t)\, dt. \tag{32}$$

Unlike the Gaussian rate (8), this does not separate. The two discrepancies enter only through the difference $\psi_k(t) - \eta_k(x)$, and they enter it multiplied by λ_k, which itself depends on both.

Proposition 11 (Local confounding of baseline and coefficient discrepancy)**.** *Suppose ψ_k and η_k are small, of order ϵ. Then*

$$I_k^{\mathrm{cox}} = \tfrac{1}{2}\,\mathbb{E}\int_0^{\bar{t}} \big\{\psi_k(t) - \eta_k(x)\big\}^2 \lambda_0(t \mid x)\, Y(t)\, dt \;+\; O(\epsilon^3), \tag{33}$$

so to second order the rate depends on $(\psi_k, \boldsymbol{\delta}^{(k)})$ only through the scalar field $\psi_k(t) - x^\top\boldsymbol{\delta}^{(k)}$. In particular, if $x^\top\boldsymbol{\delta}^{(k)}$ is constant in x on the support of the covariate distribution, then any baseline discrepancy $\psi_k(t) \equiv c$ with $c = x^\top\boldsymbol{\delta}^{(k)}$ gives $I_k^{\mathrm{cox}} = O(\epsilon^3)$: a source with a shifted baseline hazard is locally indistinguishable from one with a shifted linear predictor.

The proof is a second-order expansion of $u\log u - u + 1$ at $u = 1$ with $u = \lambda_k/\lambda_0 = \exp\{\psi_k - \eta_k\}$. The consequence is that selection in the Cox setting is identified only through the way the two discrepancies vary, ψ_k in t and η_k in x, and that a source whose baseline shift happens to mimic a shift in the linear predictor is invisible to the criterion. Whether it should be invisible is a substantive question and not a statistical one: for a prognostic model the two are interchangeable, for an etiological claim they are not.

This matters for the prior. The matrix-normal specification places independent structure on the spline coefficients and on $\boldsymbol{\delta}^{(k)}$, so it assigns them separate scales, whereas (33) says the likelihood constrains only their difference. The row covariance Ω of the matrix-normal prior therefore silently determines how a given total discrepancy is attributed between the two, and the posterior on γ_k inherits that attribution. The source material acknowledges the modelling issue in applied terms, noting that clinically similar endpoints may not be comparable across disease sites, which is the same point reached from the other direction. We regard a proper treatment as open. It cannot be settled by simulation alone at the dimensions currently feasible under Hamiltonian Monte Carlo, which is one more argument for the generative computation of Section 9.

10.3 Recommendations

The practical recommendations are four. Give each excluded source its own coefficient vector rather than pooling the excluded block, since the pooled specification is not consistent for $\mathcal{A}$, the failure does not diminish with sample size, and it costs a factor of three at small n_0. Do not expect source selection to work when $n_k < p$, since below that threshold the comparison is dominated by estimation savings common to all sources. Keep the contrast prior tight, since at small pool sizes it is the only thing supplying identification, and report sensitivity to it, since the inclusion probabilities then depend on

it more than on the data. And benchmark at the target sample size at which the method will be deployed, since the four standard metrics cannot see a selection failure at the sample sizes currently used.

What we have not shown. Propositions 4 and 5 are stated for fixed p with $n \to \infty$ and with known variance; the joint regime $p, n, K \to \infty$ is open and is the one that matters for genomics. The enumeration uses ridge priors in place of the horseshoe, so it isolates the structural effects at the cost of not testing whether adaptive shrinkage repairs any of them. Six replicates is few, and Section 6.3 varies n_0 at a single contamination level and a single contrast scale. Proposition 9 supplies the threshold exactly and matches the simulated excess to 2%, but it is a statement about the posterior mean under a known design and known variance, not a bound. Section 5.1 relaxes the known variance and finds the pooled specification degrades further rather than recovering, but its hierarchy is simpler than the published one and its inversion result rests on three replicates. The remaining untested form of adaptivity is local: whether coordinate-specific shrinkage, the horseshoe proper rather than a ridge, repairs Proposition 4. We doubt it, since the mechanism in Proposition 4 concerns which block explains a source rather than how its coefficients are shrunk, but testing it requires replacing the enumeration by a sampler, which reintroduces exactly the mixing question the enumeration was designed to avoid. The sample size bound (17) treats the flips as independent, which they are not, and should be read as an order-of-magnitude statement. The Sanov formulation of Section 3 is at present a definition and a rate calculation rather than a procedure; turning Definition 1 into an estimator, presumably through a predictive discrepancy between the target posterior predictive and each source's, is the piece we would write next. Finally, the covariate-shift term in (8) is untouched by the contrast parameterization, and separating a coefficient discrepancy from a design discrepancy is, in our view, the substantive open problem in this area.

A Proofs

Throughout, $\Sigma_k = \mathbb{E}[x^{(k)}(x^{(k)})^\top]$, the noise variance σ^2 is common and known, and $\hat{\Sigma}_k = n_k^{-1} X^{(k)\top} X^{(k)} \to \Sigma_k$ almost surely. We write $m(\cdot \mid \boldsymbol{\gamma})$ for the marginal likelihood under configuration $\boldsymbol{\gamma}$ and suppress σ^2.

A.1 Propositions 2 and 3

Under (4) the conditional laws of y given x under P_k and P_0 are $\mathcal{N}(x^\top \mathbf{w}^{(k)}, \sigma^2)$ and $\mathcal{N}(x^\top \boldsymbol{\beta}, \sigma^2)$. For two Gaussians with common variance the divergence is half the squared standardized mean difference, so

$$\mathrm{KL}\big(P^{(k)}_{y|x} \,\|\, P^{(0)}_{y|x}\big) = \frac{\big(x^\top(\boldsymbol{\beta} - \mathbf{w}^{(k)})\big)^2}{2\sigma^2} = \frac{(x^\top \boldsymbol{\delta}^{(k)})^2}{2\sigma^2}. \tag{34}$$

The chain rule for relative entropy on the joint law of (x, y) gives $\mathrm{KL}(P_k \| P_0) = \mathrm{KL}(P_{x^{(k)}} \| P_{x^{(0)}}) + \mathbb{E}_{P_{x^{(k)}}} \mathrm{KL}(P^{(k)}_{y|x} \| P^{(0)}_{y|x})$, and $\mathbb{E}_{P_{x^{(k)}}}[(x^\top \boldsymbol{\delta}^{(k)})^2] = (\boldsymbol{\delta}^{(k)})^\top \Sigma_k \boldsymbol{\delta}^{(k)}$. Equation (8) follows. □

For Proposition 3, with $P_{x^{(k)}} = P_{x^{(0)}}$ the covariate term vanishes and $I_k = (2\sigma^2)^{-1} (\boldsymbol{\delta}^{(k)})^\top \Sigma \boldsymbol{\delta}^{(k)}$. The Rayleigh quotient bounds give $\lambda_{\min}(\Sigma)\|\boldsymbol{\delta}^{(k)}\|_2^2 \le (\boldsymbol{\delta}^{(k)})^\top \Sigma \boldsymbol{\delta}^{(k)} \le \lambda_{\max}(\Sigma)\|\boldsymbol{\delta}^{(k)}\|_2^2$, so $\|\boldsymbol{\delta}^{(k)}\|_2 \le t$ implies $I_k \le \lambda_{\max}(\Sigma) t^2/(2\sigma^2)$ and $I_k \le \lambda_{\min}(\Sigma) t^2/(2\sigma^2)$ implies $\|\boldsymbol{\delta}^{(k)}\|_2 \le t$. That $I_k = 0$ characterizes $\mathcal{A}^\star$ follows from Definition 1 and the fact that relative entropy vanishes only for coinciding laws, together with the full rank of Σ, which makes $\boldsymbol{\delta}^{(k)} \mapsto (\boldsymbol{\delta}^{(k)})^\top \Sigma \boldsymbol{\delta}^{(k)}$ positive definite. □

A.2 Proposition 4

Fix p and let $n_k = n$ for all k. Write $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ for the configurations that agree except that $\gamma_k = 1$ and $\gamma'_k = 0$. By Laplace approximation, valid under the stated regularity conditions since the models are

regular exponential families with fixed dimension,

$$\log m(y \mid \boldsymbol{\gamma}) = \ell_n(\hat{\theta}_\gamma) - \tfrac{d_\gamma}{2} \log n + O_p(1), \tag{35}$$

where d_γ is the dimension. Under (3)–(5) the two configurations have the same dimension, $2p$ for the informative block and p for the nuisance block, whenever both blocks are nonempty under both configurations, so the dimension penalties cancel and $\log \mathrm{BF} = \ell_n(\hat{\theta}_\gamma) - \ell_n(\hat{\theta}_{\gamma'}) + O_p(1)$.

By the law of large numbers, $n^{-1}\ell_n(\hat{\theta}_\gamma)$ converges to the negative of the aggregate divergence between the true source laws and the best fit within the model. Under $\boldsymbol{\gamma}$, source k is fitted by the anchor $\mathbf{w}^\star(\mathcal{A}_\gamma)$ shared with the other included sources; under $\boldsymbol{\gamma}'$ it is fitted by $\mathbf{v}^\star(\bar{\mathcal{A}}_{\gamma'})$ shared with the other excluded sources. Collecting terms and cancelling the contributions of the studies whose block membership is unchanged, up to the second-order effect of moving one study on the two pooled minimizers,

$$\frac{1}{n} \log \mathrm{BF} \longrightarrow \underbrace{\frac{(\mathbf{w}^{(k)} - \mathbf{v}^\star)^\top \Sigma (\mathbf{w}^{(k)} - \mathbf{v}^\star)}{2\sigma^2}}_{J_k^{\mathrm{out}}} - \underbrace{\frac{(\mathbf{w}^{(k)} - \mathbf{w}^\star)^\top \Sigma (\mathbf{w}^{(k)} - \mathbf{w}^\star)}{2\sigma^2}}_{I_k^{\mathrm{in}}}. \tag{36}$$

If the true excluded coefficients are not all equal then $\mathbf{v}^\star$ differs from $\mathbf{w}^{(k)}$ for at least one $k \in \bar{\mathcal{A}}$, so $J_k^{\mathrm{out}} > 0$ for that k. Neither limit involves n, so the sign of the difference is determined by the composition of the two blocks. In particular there exist configurations of true source coefficients for which the limit is positive at a $k \notin \mathcal{A}$, and the posterior then concentrates on a configuration including k for all large n. □

Remark 12. The proposition is stated for fixed p because that is where the Laplace expansion (35) is uncontroversial. The enumeration in Section 5 operates at $p = 200$ with n_k between 150 and 600, well outside that regime, and exhibits the predicted behaviour, which we regard as evidence that the mechanism survives. We do not claim to have proved it there.

A.3 Proposition 5

Under the source-specific nuisance specification the excluded model is correctly specified: $y^{(k)} \sim \mathcal{N}(X^{(k)}\mathbf{v}_k, \sigma^2 I)$ with $\mathbf{v}_k$ free and the true $\mathbf{w}^{(k)}$ in the support of the prior. Hence $J_k^{\mathrm{out}} = 0$. The dimension comparison is now nontrivial, since under $\gamma_k = 1$ source k contributes no new parameters while under $\gamma_k = 0$ it contributes p. Applying (35) with $d_\gamma - d_{\gamma'} = -p$,

$$\log \mathrm{BF} = \{\ell_n(\hat{\theta}_\gamma) - \ell_n(\hat{\theta}_{\gamma'})\} + \tfrac{p}{2} \log \frac{n_k}{2\pi} + O_p(1), \tag{37}$$

and the bracket converges to $-n_k I_k$ by the argument of the previous proof with $J_k^{\mathrm{out}} = 0$. Dividing by n_k gives the statement. Consistency for $\mathcal{A}$ requires $n_k I_k$ to dominate $\frac{p}{2} \log n_k$, that is $n_k/(p \log n_k) \to \infty$ for fixed I_k, which is implied by $n_k/p \to \infty$ up to the logarithmic factor. □

A.4 Proposition 6

Take $\mathcal{A}_\gamma = \{k\}$ and reparameterize the informative block from $(\mathbf{w}, \boldsymbol{\delta})$ to $(\mathbf{w}, \boldsymbol{\beta})$ with $\boldsymbol{\beta} = \mathbf{w} + \boldsymbol{\delta}$, a linear bijection with unit Jacobian. Under $\boldsymbol{\delta} \sim \mathcal{N}(0, \sigma^2 g_\delta I)$ the induced conditional prior is $\boldsymbol{\beta} \mid \mathbf{w} \sim \mathcal{N}(\mathbf{w}, \sigma^2 g_\delta I)$. The joint marginal likelihood is

$$m(y \mid \boldsymbol{\gamma}) = \int \Big[\int p(y^{(0)} \mid \boldsymbol{\beta})\, \mathcal{N}(\boldsymbol{\beta} \mid \mathbf{w}, \sigma^2 g_\delta I)\, d\boldsymbol{\beta} \Big] p(y^{(k)} \mid \mathbf{w})\, \pi(\mathbf{w})\, d\mathbf{w} \times \prod_{j \neq k} c_j. \tag{38}$$

As $g_\delta \to \infty$ the inner Gaussian converges to Lebesgue measure on $\mathbb{R}^p$ in the sense that, for $X^{(0)}$ of full row rank and $y^{(0)}$ fixed, the inner integral is $(2\pi\sigma^2 g_\delta)^{-p/2}$ times a quantity converging to $\int p(y^{(0)} \mid \boldsymbol{\beta})\, d\boldsymbol{\beta}$, which is finite and free of $\mathbf{w}$. Hence the $\mathbf{w}$ integral factorizes and $m(y \mid \boldsymbol{\gamma})$ depends on the target only through a multiplicative constant common to all $\boldsymbol{\gamma}$ with $|\mathcal{A}_\gamma| = 1$. The Bayes factor against $\gamma_k = 0$ therefore depends on $\mathcal{D}_k$ alone. □

Remark 13. The proposition is an exact statement about the limit. At finite g_δ the target contributes, and the size of the contribution is what Figure 2 measures: the evidence gap at $|\mathcal{A}_\gamma| = 1$ falls from 34.2 nats at $g_\delta = 0.01$ to 0.40 at $g_\delta = 1$, and the proposition says it continues to zero. This is why a truncation that bounds the contrast scale below the anchor scale is an identification device.

A.5 Proposition 9 and Corollary 10

Write $A = S_\mathcal{A} + S_0 + g_w^{-1} I$, $B = S_0$, $C = S_0 + g_\delta^{-1} I$, so that $\Lambda = \left(\begin{smallmatrix} A & B \\ B & C \end{smallmatrix}\right)$ is the posterior precision of $(\mathbf{w}, \boldsymbol{\delta})$. Under the true model source k has coefficient $\mathbf{w}^{(k)} = \boldsymbol{\beta} - \boldsymbol{\delta}^{(k)}$, so taking expectations of the normal equations,

$$\mathbb{E}\big[Z_\gamma^\top y\big] = \begin{pmatrix} \sum_k X^{(k)\top} X^{(k)} \mathbf{w}^{(k)} + S_0 \boldsymbol{\beta} \\ S_0 \boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} (S_\mathcal{A} + S_0)\boldsymbol{\beta} - M \\ S_0 \boldsymbol{\beta} \end{pmatrix}, \tag{39}$$

with M as in (22). The key step is to notice that $\Lambda(\boldsymbol{\beta}^\top, 0)^\top = ((S_\mathcal{A} + S_0 + g_w^{-1} I)\boldsymbol{\beta},\ S_0 \boldsymbol{\beta})^\top$, so

$$\mathbb{E}\big[Z_\gamma^\top y\big] = \Lambda \begin{pmatrix} \boldsymbol{\beta} \\ 0 \end{pmatrix} - \begin{pmatrix} M + g_w^{-1} \boldsymbol{\beta} \\ 0 \end{pmatrix}, \tag{40}$$

whence $\mathbb{E}[(\hat{\mathbf{w}}, \hat{\boldsymbol{\delta}})] = (\boldsymbol{\beta}, 0) - \Lambda^{-1}(M + g_w^{-1}\boldsymbol{\beta},\ 0)$ and

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] - \boldsymbol{\beta} = \mathbb{E}[\hat{\mathbf{w}} + \hat{\boldsymbol{\delta}}] - \boldsymbol{\beta} = -\big(V_{ww} + V_{\delta w}\big)\big(M + g_w^{-1}\boldsymbol{\beta}\big), \tag{41}$$

where $V = \Lambda^{-1}$. By the block inverse, $V_{ww} = (A - BC^{-1}B)^{-1}$ and $V_{\delta w} = -C^{-1} B V_{ww}$, so

$$V_{ww} + V_{\delta w} = \big(I - C^{-1} S_0\big) V_{ww} = C^{-1}\big(C - S_0\big) V_{ww} = g_\delta^{-1} C^{-1} V_{ww}, \tag{42}$$

using $C - S_0 = g_\delta^{-1} I$. This is (23).

For (24), set $g_w \to \infty$ and $S_0 = n_0 I$, so $C = (n_0 + g_\delta^{-1}) I$ and

$$A - S_0 C^{-1} S_0 = S_\mathcal{A} + S_0 - S_0 C^{-1} S_0 = S_\mathcal{A} + g_\delta^{-1} S_0 C^{-1} = S_\mathcal{A} + \frac{n_0}{1 + n_0 g_\delta} I. \tag{43}$$

Multiplying by $g_\delta^{-1} C^{-1} = \{g_\delta(n_0 + g_\delta^{-1})\}^{-1} I = (1 + n_0 g_\delta)^{-1} I$ gives the stated form. The corollary follows by putting $S_\mathcal{A} = |\mathcal{A}_\gamma| n_k I$, so that $M = n_k \Delta$, and dropping n_0/ζ against $|\mathcal{A}_\gamma| n_k$. □

Remark 14. The two simplifications are not equally innocent. Replacing $S_\mathcal{A}$ by its mean is controlled by the Marchenko-Pastur ratio $p/(|\mathcal{A}_\gamma| n_k)$ and is mild when many sources are pooled. Replacing M by $n_k \Delta$ is not: the discarded term $\sum_k (X^{(k)\top} X^{(k)} - n_k I)\boldsymbol{\delta}^{(k)}$ has squared norm of order $p\, n_k \sum_k \|\boldsymbol{\delta}^{(k)}\|_2^2$, which is comparable to $n_k^2 \|\Delta\|_2^2$ whenever $p \gtrsim n_k$, and the two are not orthogonal. Section 6.4 reports a factor of fourteen at $p/n_k = 0.67$.

B The enumeration

We record the algebra used in Section 5. Under configuration $\boldsymbol{\gamma}$, stack the informative block parameters as $\theta = (\mathbf{w}^\top, \boldsymbol{\delta}^\top)^\top \in \mathbb{R}^{2p}$ with design

$$Z_\gamma = \begin{pmatrix} X^{(\mathcal{A}_\gamma)} & 0 \\ X^{(0)} & X^{(0)} \end{pmatrix}, \qquad \theta \sim \mathcal{N}\big(0, \sigma^2 G\big), \quad G = \operatorname{diag}(g_w I_p, g_\delta I_p). \tag{44}$$

With $\Lambda_\gamma = Z_\gamma^\top Z_\gamma + G^{-1}$ and $\mu_\gamma = \Lambda_\gamma^{-1} Z_\gamma^\top y$, the informative block contributes

$$-\tfrac{1}{2}\Big\{\log|\Lambda_\gamma| + p\log g_w + p\log g_\delta + \sigma^{-2}\big(y^\top y - y^\top Z_\gamma \mu_\gamma\big)\Big\} \tag{45}$$

to $\log m(y \mid \boldsymbol{\gamma})$, where y collects the target and included sources. Because $Z_\gamma^\top Z_\gamma$ depends on $\boldsymbol{\gamma}$ only through $\sum_{k\in\mathcal{A}_\gamma} X^{(k)\top} X^{(k)}$, the per-source cross products are computed once and accumulated. Under the source-specific nuisance specification the excluded block contributes $\sum_{k\notin\mathcal{A}_\gamma} c_k$ with

$$c_k = -\tfrac{1}{2}\Big\{\log\big|X^{(k)\top}X^{(k)} + g_v^{-1} I_p\big| + p\log g_v + \sigma^{-2}\big(y^{(k)\top} y^{(k)} - b_k^\top S_k^{-1} b_k\big)\Big\}, \tag{46}$$

$b_k = X^{(k)\top} y^{(k)}$ and $S_k = X^{(k)\top} X^{(k)} + g_v^{-1} I_p$, so the c_k are computed once for all configurations. Under the pooled specification the excluded block requires one $p \times p$ factorization per configuration. Total cost is $O(2^K p^3)$ dominated by the $2p \times 2p$ Cholesky of Λ_γ; at $K = 10$, $p = 200$ this is a few seconds on one core.

Posterior summaries follow. The target posterior mean under $\boldsymbol{\gamma}$ is $\mathbb{E}[\boldsymbol{\beta} \mid \mathcal{D}, \boldsymbol{\gamma}] = (I_p, I_p)\mu_\gamma$ and its covariance is $(I_p, I_p)\Lambda_\gamma^{-1}(I_p, I_p)^\top$, from which Table 3 and the interval widths of Section 6 are obtained by the law of total variance across configurations weighted by $p(\boldsymbol{\gamma} \mid \mathcal{D}) \propto m(y \mid \boldsymbol{\gamma}) p(\boldsymbol{\gamma})$.

C Reproduction

All experiments are exact enumerations in a conjugate model and involve no Monte Carlo beyond the generation of synthetic data. The data generating process follows the simulation design of Jamshidian (2026): $K = 10$ sources, $p = 200$ predictors, $n_0 = 150$ target observations, six signals of size 0.5, informative sources deviating by 0.3 on two randomly chosen coordinates and noninformative sources by 0.6 on twelve, covariates independent standard normal, noise variance one. Hyperparameters are $g_w = g_v = 1$, $g_\delta = 0.1$ except where varied, and $\pi = 1/2$ except in the multiplicity comparison. Reported figures are averages over six replicates for Tables 1 and 3 and three for the interval widths, with seeds $1, \ldots, 6$.

Acknowledgements

The empirical target of this note is the doctoral dissertation of Jamshidian (2026), which we read with admiration. Nothing here should be taken as diminishing it: the propagation of selection uncertainty into target credible intervals is a real contribution, and the specification issues we identify are visible only because the framework is stated precisely enough to be probed.

References

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016). Default Bayesian analysis with global-local shrinkage priors. *Biometrika* 103, 955–969.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika* 103, 985–991.

Bondesson, L. (1992). *Generalized Gamma Convolutions and Related Classes of Distributions and Densities*. Springer, New York.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* 97, 465–480.

Jamshidian, P. M. (2026). *Bayesian Transfer Learning for High-Dimensional Regression Models via Adaptive Shrinkage*. PhD thesis, University of California, Los Angeles. https://escholarship.org/uc/item/85n604jw

Li, S., Cai, T. T., and Li, H. (2022). Transfer learning for high-dimensional linear regression: prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society B* 84, 149–173.

Polson, N. G. and Scott, J. G. (2011). Shrink globally, act locally: sparse Bayesian regularization and prediction. *Bayesian Statistics 9*, 501–538.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association* 108, 1339–1349.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics* 38, 2587–2619.

Song, Q. and Liang, F. (2023). Nearly optimal Bayesian shrinkage for high-dimensional regression. *Science China Mathematics* 66, 409–442.

Tian, Y. and Feng, Y. (2023). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association* 118, 2684–2697.