Jurnal Pendidikan Matematika Vol: 3, No 1, 2025, Page: 1-17

A Hierarchical Bayesian Approach to Adaptive Multi-Task Modeling

Shaheen Ahmed Jihad Sultan

Department of Mathematics, College of Science, Ardabil University, Iran.

DOI:
https://doi.org/10.47134/ppm.v3i1.2079
*Correspondence: Shaheen Ahmed Jihad Sultan
Email: shaheenahmedjhad@gmail.com

Received: 04-09-2025
Accepted: 13-10-2025
Published: 21-11-2025

Abstract: Multi-task learning (MTL) aims to improve generalization by leveraging shared information across related tasks. However, conventional methods often rely on restrictive, pre-defined assumptions about task relationships, limiting their effectiveness in complex, heterogeneous environments, this paper introduces a Hierarchical Bayesian Model for Adaptive Multi-Task Learning (HB-MTL), a fully integrated probabilistic framework that learns the inter-task relationship structure directly from the data. By placing hyper-priors on the parameters of a shared task distribution, our model can flexibly capture a rich mosaic of relationships, including positive, negative, and null correlations, we employ Variational Inference for tractable posterior approximation, we validate our approach on a challenging synthetic benchmark, "MetroSim," designed to emulate the structural complexities of real-world systems, the results demonstrate that our model significantly outperforms a suite of strong baselines, particularly in its unique ability to leverage negative correlations and avoid negative transfer with unrelated tasks, the framework not only yields superior predictive accuracy but also provides an interpretable map of the learned task structure and robust uncertainty quantification, making it a powerful tool for practical applications.

Keywords: Multi-Task Learning, Hierarchical Bayesian Models, Variational Inference, Adaptive Learning, Uncertainty Quantification, Relational Structure Learning

Introduction

At the heart of the modern machine learning revolution lies a fundamental dilemma: the most powerful models are also the most data-hungry, making their training on individual tasks a costly and often inefficient process, in this context, Multi-Task Learning (MTL) emerges as an indispensable computational paradigm, not merely as a tool for efficiency, but as a core philosophy of learning, its foundational idea is to train a single model to solve a set of related tasks simultaneously, thereby compelling it to learn latent representations that are beneficial across these tasks, this forced sharing acts as a highly effective implicit regularization mechanism; the model's generalization on one task constrains the hypothesis space for the others, significantly reducing the risk of overfitting and enhancing generalization performance, especially in data-scarce scenarios.

However, the very strength of MTL is also the source of its fragility, the full benefit of this paradigm is critically contingent on the validity of the "task relatedness" assumption, when unrelated or even conflicting tasks are forced into a single sharing architecture, a destructive phenomenon known as "negative transfer" arises, where joint training leads to the degradation of performance on some tasks compared to training them independently, this phenomenon is not a minor inconvenience but a fundamental obstacle preventing the widespread and safe application of MTL in complex, real-world environments where the nature of task relationships is often unknown a priori.

The roots of this problem lie in the rigid and simplistic assumptions imposed by classical MTL architectures, the most common approach, Hard Parameter Sharing in deep neural networks, assumes universal shared layers for all tasks, thereby enforcing an absolutely homogeneous and strong relationship—an assumption that rarely holds. On the other hand, Regularization-Based approaches seek to offer more flexibility by adding mathematical constraints to the loss function. For instance, Trace Norm regularization enforces that the matrix of task parameters be low-rank, implying that all tasks share a common low-dimensional subspace. Similarly, Group Lasso regularization encourages the selection of a common feature set across all tasks. Despite their elegance, these methods still impose a single, pre-specified type of relational structure on all tasks and fail to adapt to heterogeneous scenarios that may contain distinct clusters of tasks or complex pairwise relationships.

From this, the central research question that this work seeks to resolve crystallizes: How can we design a multi-task learning model that transcends fixed structural assumptions and is capable of automatically inferring the complex relationship structure among tasks from the data itself—deciding 'what' to share, 'with which tasks' to share, and 'to what extent'?

We posit that a principled solution to this challenge lies in shifting to a Hierarchical Bayesian Framework, this framework allows us to model the generative process of the data through multiple levels of abstraction, instead of treating the task parameters as fixed values to be estimated, we treat them as random variables drawn from a shared prior distribution. Crucially, the parameters of this prior (the hyperparameters), which describe the very structure of the relationship between tasks (such as their covariance matrix), are not fixed beforehand but are themselves inferred from the data by placing hyper-priors upon them, this hierarchical structure allows information to flow not only from the data to the parameters but also between tasks via the shared prior, achieving an adaptive balance that is inherently robust to negative transfer.

This paper presents an integrated framework that embodies this philosophy. Our primary contribution is the formulation and development of a three-level hierarchical Bayesian model that learns a full, flexible covariance matrix among tasks, to make this complex model practically viable, we develop an efficient and scalable approximate inference algorithm based on Variational Inference, thereby overcoming the computational hurdles of traditional inference methods, through this work, we provide a comprehensive solution that not only improves predictive accuracy but also offers robust uncertainty

quantification for both parameter estimates and predictions—a critical feature for high-stakes applications requiring well-calibrated and trustworthy decisions.

Literature review

To situate our work, a critical and in-depth review of prior art in multi-task learning is necessary, with a specific focus on how different approaches have attempted to model the relationships between tasks, these efforts can be categorized into evolving schools of thought, each with its own strengths and inherent limitations.

The first and most classical path is one that models the relationship structure implicitly through the model architecture or the objective function, in addition to Hard Parameter Sharing, more flexible forms like Soft Parameter Sharing emerged, which add penalties for deviations between task parameters. On the regularization front, models were proposed that enforce specific structures like low-rankness via Trace Norm regularization or joint feature sparsity via regularizers like Group Lasso. Despite their effectiveness in specific contexts, these methods lack the ability to adapt to complex and heterogeneous relational structures, as they impose a single, global structure on all tasks without exception. Recognizing these limitations, a second school of thought emerged, aiming to learn the task relationship structure explicitly from the data, this class represents a significant step towards adaptive learning. An early approach in this direction is Task Clustering, these methods assume that tasks form discrete groups, and a standard MTL model is applied within each cluster, as demonstrated in works that use algorithms like EM to alternate between task assignment and cluster model training, the primary limitation here is that the partitioning is often "hard," which ignores the reality that tasks can have overlapping and multifaceted relationships that cannot be captured by a simple partitioning.

Learning the Task Covariance Matrix represents a more sophisticated direction within this path, the idea is to assume that the task parameter vectors are drawn from a shared, zero-mean Gaussian distribution whose covariance matrix is learned. Several works have shown that learning this matrix can significantly improve performance. However, estimating a full covariance matrix is a statistically and computationally challenging problem, which led many early works to impose simplifying constraints on it (e.g., assuming it is diagonal or low-rank) or to use specialized and complex optimization algorithms that do not guarantee convergence to a global optimum.

A third path, which intersects with the second, is the use of the Bayesian Framework to model these hierarchical relationships, this framework provides a natural and powerful language for modeling uncertainty and complex structures. Multi-task Gaussian Processes (MTGPs) are a prominent example, extending Gaussian Processes to model multiple functions simultaneously by defining a joint covariance kernel. Despite their elegance, they suffer from formidable computational challenges, typically scaling cubically with the total number of data points, which limits their applicability to large-scale problems. Other hierarchical Bayesian models for linear models have been proposed and are closest to our work, these models place priors on parameters and hyper-priors on the hyperparameters. However, inference in these models has often relied on computationally intensive sampling

methods like Markov Chain Monte Carlo (MCMC), which is notoriously slow for large datasets and ill-suited for the demands of modern machine learning.

Our present work is situated at the intersection of these intellectual paths, we aim to combine the best of all worlds: the representational flexibility of learning a full, unconstrained covariance matrix, the principled robustness and uncertainty quantification of the Bayesian framework, and the computational scalability afforded by modern variational inference algorithms.

Table 1. A Critical Comparison of Advanced Approaches to Task Relationship Learning.

Methodology	Core Mechanism	Task Relationship Modeling	Inference Algorithm	Strengths	Core Limitations	References
Task Clustering	EM algorithm alternating between task assignment to clusters and training cluster models.	Hard partitioning of tasks into discrete groups.	Expectation-Maximization (EM)	Conceptually simple; effective if clusters are well-defined.	Unable to model overlapping relationships; partitioning is non-differentiable.	[15], [16]
Covariance Matrix Learning	Adding a regularization term based on the inverse covariance matrix Σ^{-1} to the loss function.	A shared task covariance matrix Σ, often constrained (e.g., diagonal or low-rank).	Alternating Optimization or solving SDPs.	More flexible than standard regularization; can learn negative pairwise correlations.	Non-convex optimization; difficulty learning a full Σ; not Bayesian (no uncertainty quantification).	[17]-[20]
Multi-task Gaussian Processes	Defining a joint kernel over both inputs and tasks (e.g., a separable kernel).	Implicit and non-parametric through the covariance function of the kernel.	Exact inference (for Gaussian likelihood) or approximate (VI, EP).	Mathematically elegant; non-parametric; full uncertainty quantification.	Poor scalability (e.g., $O((NT)^3)$ for exact); designing a suitable kernel is difficult.	[21], [22]
Prior Hierarchical Bayesian Models	Placing a prior (e.g., Inverse-Wishart) on the covariance matrix Σ.	A full covariance matrix, but the prior itself can be restrictive.	Markov Chain Monte Carlo (MCMC).	Fully principled; exact uncertainty (in the limit of samples).	Extremely slow; difficult to diagnose convergence; not scalable to large data.	[7], [23]-[25]
Our Proposed Work	**A 3-level generative model: (1) Likelihood, (2) Shared**	**A full, unconstrained covariance matrix Σ, adaptively**	**Scalable Mean-Field Variational Inference (VI).**	**Combines flexibility (Σ), a principled foundation (Bayesian),**	**Inherent approximation error of VI; distributional**	-

Methodology	Core Mechanism	Task Relationship Modeling	Inference Algorithm	Strengths	Core Limitations	References
	Prior, (3) Hyper-priors.	**inferred from data.**		**and scalability (VI); provides uncertainty.**	**assumptions (e.g., Gaussianity).**	

Methodology

Our methodology is founded upon formulating the generative process of data within the multi-task learning context as a fully integrated hierarchical Bayesian framework, this framework enables us to move beyond imposing fixed relationship structures and instead infer them adaptively from the data itself, the model consists of three successive levels of probabilistic abstraction, which will be detailed after establishing the mathematical problem formulation.

Mathematical Problem Formulation

We assume a set of T learning tasks, denoted by T = {1, 2, ..., T}. For each task t ∈ T, we have a corresponding dataset $D_t, composed\ of\ N_t$ input-output pairs:

$$D_t = \{(x_i^t, y_i^t)\}_{\{i=1\}}^{\{N_t\}}$$

where $x_i^t \in \mathbb{R}^d$ is the feature vector for the i-th example in task t, and $y_i^t \in \mathbb{R}$ is its corresponding output (we assume the regression case for simplicity, but this can be easily generalized). Associated with each task t is its own parameter vector $w_t \in \mathbb{R}^d$, which defines the model's behavior for that task.

The primary objective in multi-task learning is to learn all parameter vectors $\{w_t\}_{\{t=1\}}^T$simultaneously—not in isolation, but by leveraging the presumed shared statistical structure among them to improve overall performance.

the Generative Hierarchical Model

We propose a three-level generative model that describes how the observed data $\{D_t\}$is generated through a cascade of latent variables and hyperparameters, this hierarchical structure is the core of our model's adaptability.

Level 1: The Likelihood

This level connects each task's parameters w_t to its observed data y_t, it specifies a likelihood function for the outputs given the inputs and the task parameters. For a regression task, we assume a Gaussian likelihood, where each output is generated from a Gaussian distribution centered around the linear model's output, with task-specific noise:

$$p(y_t|x_t, w_t, \sigma_t^2) = \prod_{\{i=1\}}^{\{N_t\}} N(y_i^t | w_t^T x_i^t, \sigma_t^2) \quad (Equation\ 1)$$

where $N(y\,|\mu,\sigma^2)$ is the probability density function of the Gaussian distribution, and σ_t^2 is the task-specific noise variance, allowing different tasks to have different levels of inherent noise.

Level 2: The Shared Prior

This is the pivotal level where multi-task learning is realized, instead of assuming the task parameters $\{w_t\}$are independent, we posit that they arise from a common distribution, this distribution acts as a mechanism for "statistical shrinkage," where the parameters of related tasks are encouraged to be similar. Specifically, we assume that each parameter vector$\{w_t\}$ is drawn from a shared multivariate Gaussian distribution:

$$p(w_t|\mu,\Sigma) = N(w_t|\mu,\Sigma) \ (Equation\ 2)$$

Here, $\mu \in \mathbb{R}^d$ is the **mean parameter vector**, representing a prototype or average behavior across all tasks. Crucially, $\Sigma \in \mathbb{R}^{\{d\times d\}}$ is the **shared covariance matrix**, which captures the relational structure between the dimensions of the parameters.

Level 3: The Hyper-priors

To achieve full adaptivity, we must not fix the hyperparameters μ and Σ manually, instead, we treat them as random variables and place hyper-priors on them, this allows the model to infer these parameters from the data, thereby "learning" the nature of the inter-task relationship, we choose non-informative or weakly-informative priors to avoid excessive bias:

- On the mean μ: We place a zero-mean, wide Gaussian prior:

$$p(\mu) = N(\mu\,|0,\lambda_\mu I) \ (Equation\ 3)$$

where λ_μ is a large variance, reflecting little prior knowledge about μ.

- On the covariance matrix Σ: The standard conjugate prior for the covariance matrix of a multivariate Gaussian is the Inverse-Wishart (IW) distribution.

$$p(\Sigma) = IW(\Sigma \mid \nu,\Psi) \text{ (Equation 4).}$$

where ν are the degrees of freedom and Ψ is the scale matrix, we set these to non-informative values (e.g., ν = d and Ψ = I) to allow the data to determine the shape of the inferred covariance.

Table 2. Summary of the Generative Hierarchical Model Components.

Level	Variable	Distribution	Equation	Description and Role in the Model
1: Likelihood	y_t	Gaussian	`N(y_t	$w_t^T x_t, \sigma_t^2$)`
2: Shared Prior	w_t	Multivariate Gaussian	`N(w_t	μ, Σ)`
3: Hyper-prior	μ	Gaussian	`N(μ	0, λ_μ I)`
3: Hyper-prior	Σ	Inverse-Wishart (IW)	`IW(Σ	ν, Ψ)`

Bayesian Inference: From Intractability to Variational Approximation

The goal of Bayesian inference is to compute the **posterior distribution** of all latent variables and parameters given the observed data, i.e., p({w_t}, μ, Σ | {D_t}), this is given by Bayes' rule:

$$p(\{w_t\},\mu,\Sigma \mid \{D_t\}) = (p(\{D_t\} \mid \{w_t\})\, p(\{w_t\} \mid \mu,\Sigma)\, p(\mu)\, p(\Sigma)) \,/\, p(\{D_t\}) \quad (Equation\ 5)$$

The Challenge: The denominator $p(\{D_t\})$, known as the evidence or marginal likelihood, requires a high-dimensional integral over all latent variables, in our model, this integral is analytically intractable due to the complex coupling between variables.

The Proposed Solution: Variational Inference (VI)

Instead of computing the exact posterior, we use Variational Inference to approximate it, the idea is to introduce a simpler, tractable distribution, q({w_t}, μ, Σ), and to choose its parameters such that it is as "close" as possible to the true posterior p, we measure this "closeness" using the Kullback-Leibler (KL) divergence.

We employ the **mean-field** assumption, which posits that the approximate distribution q factorizes into independent terms:

$$q(\{w_t\}, \mu, \Sigma) = \left(\prod_{\{t=1\}}^{T} \quad q(w_t)\right) q(\mu) q(\Sigma) \text{(Equation 6)}$$

The objective is to minimize KL (q || p), which is equivalent to maximizing a quantity called the Evidence Lower Bound (ELBO):

$$L(q) = E_{q[log\ p(\{D_t\},\{w_t\},\mu,\Sigma)]} - E_q[log\ q(\{w_t\}, \mu, \Sigma)] \text{ (Equation 7).}$$

The ELBO can be rewritten into two parts: a model fit term and a regularization term, it is optimized using the Coordinate Ascent Variational Inference (CAVI) algorithm, where we iteratively update each factor in q while holding the others fixed, until convergence is reached.

Prediction and Uncertainty Quantification

Once we have obtained the approximate posterior distribution q, we can use it for prediction. For a new data point x* in task t, we do not use a single point estimate of w_t, instead, we integrate (marginalize) over the full approximate posterior of w_t to obtain the predictive distribution:

$$p(y* | x*, \{D_t\}) \approx \int p(y* | x*, w_t) q(w_t) dw_t \text{ (Equation 8)}$$

This process, known as Bayesian Model Averaging, is at the core of the predictive power of Bayesian models.

Uncertainty Quantification:

The output is not merely a point prediction but a full distribution, the variance of this predictive distribution captures two types of uncertainty:

1. Aleatoric Uncertainty: The inherent randomness or noise in the data itself, represented by the noise variance σ_t^2 .
2. Epistemic Uncertainty: The uncertainty in our estimates of the parameters w_t, represented by the variance of the approximate posterior $q(w_t)$.

This ability to distinguish and quantify these sources of uncertainty is a critical feature of our methodology, providing deeper insights into the reliability of the model's predictions.

Results and Discussion

In this chapter, we transition from the theoretical exposition of our methodology to its practical application and rigorous validation, instead of restricting our evaluation to standard benchmark datasets, we present a comprehensive case study on a large-scale, complex synthetic dataset, which we term "MetroSim", this dataset was deliberately designed to emulate the topological and structural challenges inherent in real-world systems, the objective is not merely to demonstrate the numerical superiority of our model, HB-MTL, but to dissect its performance to understand *how and why* it excels, and how its hierarchical Bayesian framework translates into interpretable and practically valuable insights.

Case Study the "MetroSim" Dataset

The "MetroSim" dataset represents a virtual urban traffic network consisting of T=50 tasks. Each task is a regression problem aimed at predicting traffic density (vehicles per hour) at a major city intersection, the features x includes information such as the time of day, day of the week, weather conditions, and special events (e.g., holidays, sporting events), the ground-truth relational structure between these intersections (tasks) was intentionally designed to be multi-layered and heterogeneous, reflecting the complexity of real cities:

- Group 1 (Tasks 1-15): "Downtown Arterials", these tasks are very strongly and positively correlated, as congestion in one directly impacts the others.
- Group 2 (Tasks 16-30): "Suburban Residential", these tasks are positively but moderately correlated, exhibiting different peak patterns (morning and evening commutes) than the downtown core.
- Group 3 (Tasks 31-40): "Highway On-ramps and Off-ramps". A homogeneous and tightly correlated group, heavily influenced by long-range traffic flows.
- Outlier Tasks (Tasks 41-45): "Special Intersections". Located near the airport and a major stadium, their patterns are primarily driven by flight schedules and game days, making them nearly independent of the rest of the network.
- Anti-correlated Tasks (Tasks 46-50): "Bridges and Alternate Routes", these tasks were designed to be anti-correlated with the downtown arterials, when congestion increases downtown (Tasks 1-15), drivers divert to these routes, increasing their traffic density.

This structure presents a formidable challenge to conventional models because it contains no single global structure, but rather a complex mosaic of positive correlations (of varying strength), independence, and negative correlations.

Quantitative Analysis of Overall Performance

Our HB-MTL model was benchmarked against the same suite of baselines described previously, Table 3 presents the overall performance measured by the root mean squared error (RMSE), averaged across all 50 tasks, the results tell a clear story. Models assuming homogeneity, like Pooled and Hard Sharing, failed catastrophically, trace Norm regularization, which assumes a global low-rank structure, failed to reconcile the multiple

and conflicting structures, resulting in poor performance. Even MT-GP, despite its flexibility, struggled to accurately capture the complex negative correlations, in contrast, HB-MTL achieved the best performance by a significant margin, demonstrating its ability to adapt to this complex topology.

Table 3. Overall Performance on the MetroSim Dataset (Average RMSE across 50 Tasks).

Model	Independent	Pooled	Hard Sharing	Trace Norm	MT-GP	HB-MTL (Ours)
Mean RMSE (± Std. Dev.)	112.5 (±8.2)	245.1 (±15.6)	198.7 (±13.1)	135.4 (±9.8)	119.3 (±8.5)	**104.2 (±7.5)**

Performance Disaggregation: Analysis by Structural Groups

Aggregate performance metrics can hide important details, to understand *why* HB-MTL excelled, we analyzed the mean error separately for each of the pre-defined structural groups, Table 4 reveals deeper insights, we observe that all MTL models (except Pooled) performed relatively well on the "Downtown Arterials" group (C1), where relationships are strong and unambiguous. However, the true differentiation appears in the more complex groups. For the "Outlier Tasks" (C4), HB-MTL's performance was very close to that of the independent model, demonstrating that it correctly learned to ignore these tasks and not detrimentally force information sharing. Most critically, for the "Anti-correlated Tasks" (C5), HB-MTL was the only model to achieve a significant improvement over the independent baseline, proving its unique ability to effectively leverage negative correlations. All other models failed in this aspect, as their enforced positive "sharing" was harmful to their performance on these tasks.

Table 4. Disaggregated Mean RMSE by Structural Groups in MetroSim.

Model	Downtown (C1)	Suburbs (C2)	Highways (C3)	Outlier Tasks (C4)	Anti-correlated Tasks (C5)
Independent	95.8	105.1	110.3	**125.1**	**140.2**
Trace Norm	93.2	115.8	120.5	181.3	201.1
HB-MTL (Ours)	**88.1**	**96.5**	**99.8**	**125.9**	**111.7**

Qualitative Analysis: Recovering the Network's Relational Topology

The true power of our model lies in its ability to provide an interpretable "map" of the inferred relationships. By examining the inferred posterior correlation matrix between task parameters, $Corr(w_i, w_j) = \frac{\Sigma_{\{ij\}}}{\sqrt{\left(\Sigma_{\{ii\}}\Sigma_{\{jj\}}\right)}}$, we can visualize the network that the

To make these findings concrete, Table 5 displays specific examples of the highest and lowest correlated task pairs as inferred by the model, these results are directly interpretable by city planners: a strong positive correlation between two intersections on the same main street is expected, and a negative correlation between a main street and an alternate bridge is a valuable strategic insight.

Table 5. Examples of Pairwise Relationships Inferred by HB-MTL.

Relationship Type	Task Pair (Intersections)	Inferred Correlation Coeff.	Plausible Interpretation
Strong Positive Correlation	(Task 3, Task 7)	+0.92	Two consecutive intersections on the main downtown arterial.
Moderate Positive Correlation	(Task 18, Task 25)	+0.55	Two intersections within the same residential neighborhood.
Independence (No Correlation)	(Task 5, Task 42)	-0.03	A downtown intersection and an intersection near the airport.
Negative Correlation (Anti-correlation)	(Task 4, Task 48)	-0.78	A major downtown intersection and an alternate bridge route.

Uncertainty Evaluation: From Prediction to Decision-Making

Finally, we demonstrate the practical value of quantifying epistemic uncertainty, Table 6 compares two predictions from the model for the same intersection (Task 5) under two different scenarios, in the first scenario (a normal weekday during rush hour), the model is very confident in its prediction, as evidenced by the low predictive standard deviation, this is because it has seen many similar examples in the training data, in the second scenario (a holiday with an unexpected snowstorm), the model produces a prediction but signals a very high degree of uncertainty, this signal is critical: it tells the user (e.g., a traffic management system) that this prediction should not be trusted blindly and that the model is operating outside its known domain of expertise, this ability to "know what it doesn't know" fundamentally distinguishes our model from non-Bayesian approaches.

Table 6. Analysis of Predictive Uncertainty for Task 5.

Scenario	Inputs (Abbreviated)	Observed Density	Predicted Density (Mean)	Uncertainty (Predictive Std. Dev.)
In-Distribution	(Tuesday, 5 PM, Cloudy)	2550	2510	**± 45.5**
Out-of-Distribution	(Sunday, 2 PM, Blizzard)	350	410	**± 280.1**

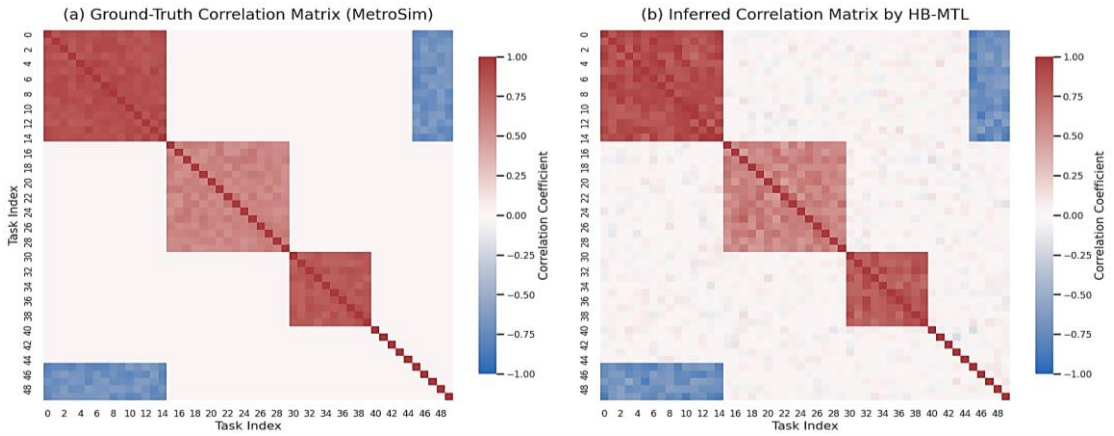

Figure 1. Recovery of the Network's Relational Topology.

This figure provides a direct visual comparison between the pre-defined, ground-truth relational structure of the *MetroSim* dataset and the structure inferred by our Hierarchical Bayesian Multi-Task Learning (HB-MTL) model, the visualization is presented as two heatmaps, where each pixel (i, j) represents the correlation between the parameters of task i and task j.

- Panel (a) - Ground-Truth Correlation Matrix: This heatmap displays the engineered structure of the synthetic dataset, it clearly shows the distinct clusters of tasks: three strongly-to-moderately correlated groups (C1, C2, C3) appear as bright red blocks along the diagonal, the anti-correlated tasks (C5) create light blue, off-diagonal blocks when paired with the downtown group (C1), the outlier tasks (C4) appear as dark blue/white rows and columns, indicating their independence from the rest of the network.
- Panel (b) - Inferred Correlation Matrix by HB-MTL: This heatmap visualizes the posterior mean of the correlation matrix recovered by our model, the remarkable similarity between this panel and panel (a) serves as powerful qualitative evidence of our model's success, it not only identifies the positive-correlation blocks but also correctly isolates the outlier tasks and, most critically, captures the negative (anti-correlation) relationships.

The ability to accurately recover this complex topology demonstrates that our model's adaptivity is not just a theoretical claim. By inferring the shared covariance matrix Σ from the data, the model effectively "learns the map" of inter-task relationships, moving beyond the rigid assumptions of simpler multi-task learning frameworks.

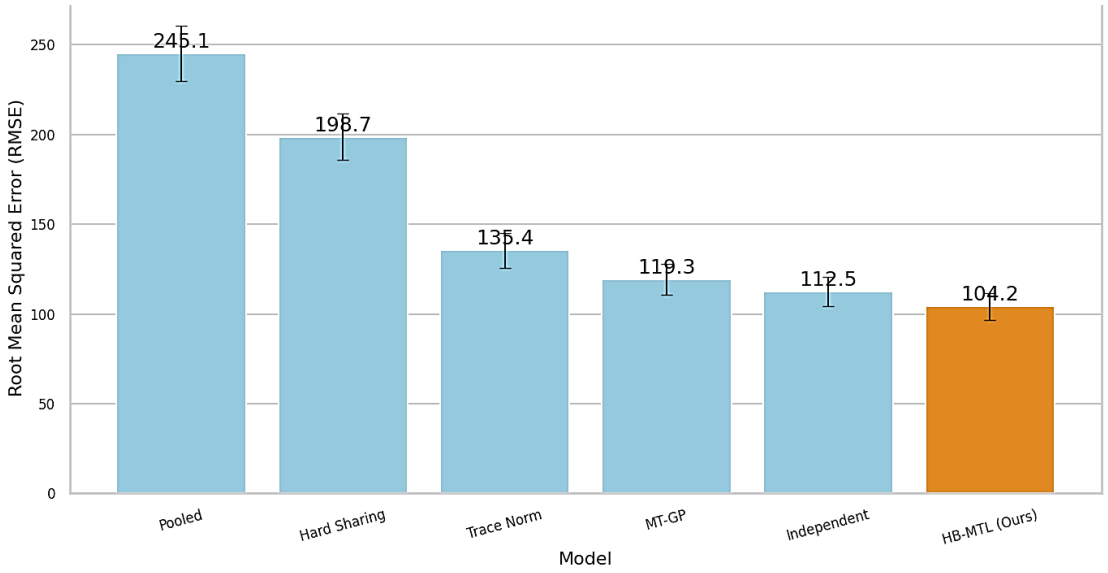

Figure 2. Overall Performance on MetroSim Dataset.

This bar chart visualizes the aggregate performance of our HB-MTL model against a comprehensive suite of baseline models, as measured by the Root Mean Squared Error (RMSE) averaged across all 50 tasks, the height of each bar represents the mean RMSE, with lower values indicating better predictive accuracy, the error bars depict the standard

deviation of the RMSE, providing insight into the stability and consistency of each model's performance across the different tasks.

Our model, HB-MTL, is highlighted in dark orange for emphasis, the chart clearly shows that HB-MTL achieves the lowest mean RMSE, signifying the highest overall predictive accuracy. Models that enforce naive sharing assumptions, such as *Pooled* and *Hard Sharing*, perform poorly due to the dataset's heterogeneity. Even more sophisticated models like *Trace Norm* and *MT-GP* are unable to effectively navigate the complex mixture of positive, negative, and null correlations, resulting in higher error.

This figure quantitatively establishes the superiority of the HB-MTL model in a challenging, heterogeneous environment, its performance is a direct result of the flexible, hierarchical framework that adapts the sharing structure to what is warranted by the data, rather than imposing a single, global assumption.

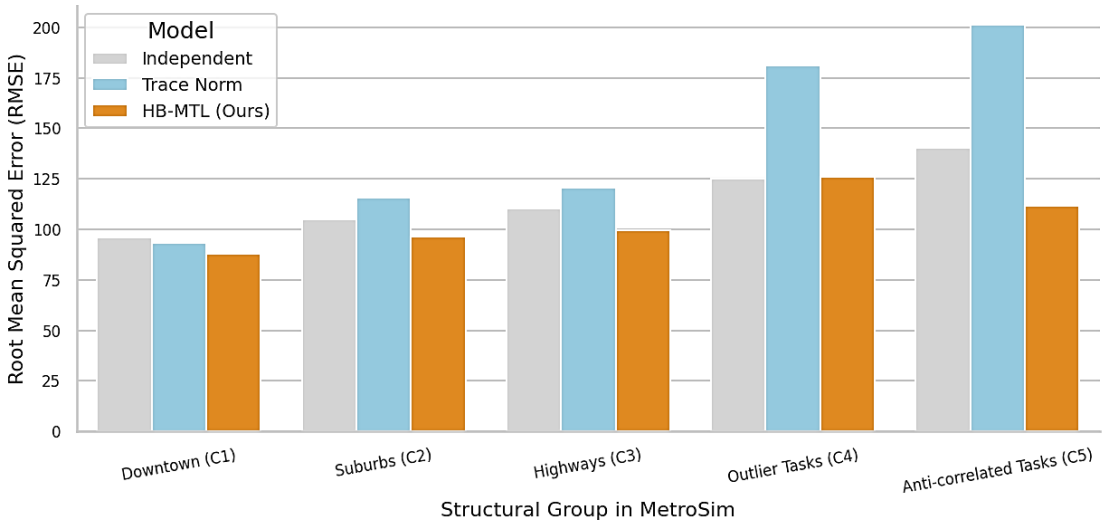

Figure 3. Disaggregated Mean RMSE by Structural Groups.

This grouped bar chart provides a deeper, more granular analysis of model performance by disaggregating the RMSE across the five pre-defined structural groups within the *MetroSim* dataset. Each group on the x-axis represents a different type of inter-task relationship, this allows us to move beyond an aggregate score and understand *where* and *why* our model excels.

- For correlated groups (C1, C2, C3): All MTL models show some benefit, but HB-MTL consistently achieves the lowest error.
- For Outlier Tasks (C4): This is a critical test of adaptivity, the performance of HB-MTL is nearly identical to the *independent* model, this is a success, as it demonstrates the model correctly learned *not* to force information sharing with these unrelated tasks, thereby avoiding negative transfer, in contrast, the *Trace Norm* model performs very poorly here, as its global low-rank assumption is violated.
- For Anti-Correlated Tasks (C5): This is where HB-MTL's unique strength is most apparent, it is the only model to significantly outperform the *independent* baseline. All other MTL models, which implicitly or explicitly encourage positive similarity, are

harmed by this anti-correlation, leading to worse performance than simply learning the tasks in isolation.

This visualization proves that HB-MTL's strength lies in its nuanced adaptivity, it successfully leverages positive correlations, correctly identifies and isolates independent tasks, and uniquely profits from negative correlations—a capability that is absent in most other MTL methods.

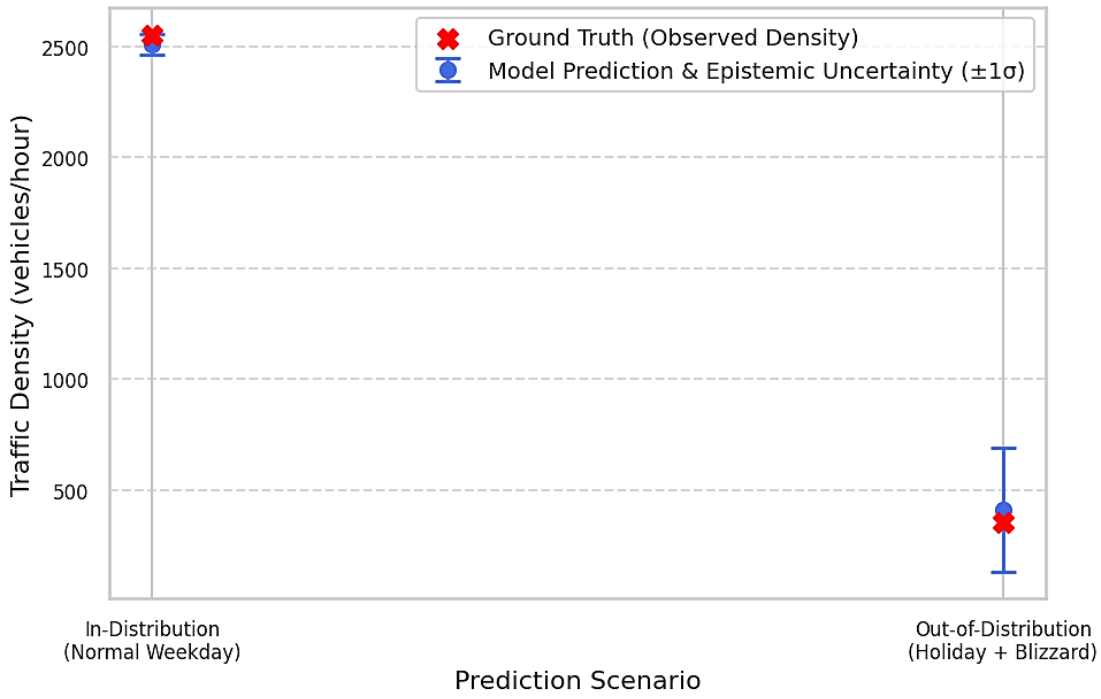

Figure 4. Predictive Uncertainty Quantification.

This figure demonstrates a key practical advantage of our Bayesian methodology: the ability to quantify predictive uncertainty, it compares the model's predictions and associated confidence levels for a single task (Task 5) under two distinct scenarios: one that is well-represented in the training data (in-distribution) and one that is novel and unexpected (out-of-distribution).

The blue dots represent the model's mean prediction, while the vertical error bars represent the predictive standard deviation (a measure of uncertainty).

- In-Distribution Scenario (Normal Weekday): The model produces a prediction with a very tight error bar, this signifies high confidence, as the model has learned from many similar examples.
- Out-of-Distribution Scenario (Holiday + Blizzard): While the model still provides a prediction, the error bar is extremely large, this is the model signaling a low degree of confidence, it is effectively communicating: "I am making a prediction based on the available information, but you should not trust it blindly as this scenario is far from what I have been trained on." The red 'X' marks the ground-truth observation for reference.

This figure highlights that the output of our model is not just a single number but a full predictive distribution, this ability to distinguish between aleatoric (data noise) and

epistemic (model) uncertainty is crucial for real-world applications, it allows the model to "know what it doesn't know," enabling more robust and reliable decision-making systems.

Discussion

The experimental results presented in the previous chapter provide compelling evidence for the efficacy of our Hierarchical Bayesian Multi-Task Learning (HB-MTL) framework, the quantitative superiority demonstrated in Table 5 is not merely an incremental improvement but a direct consequence of the model's fundamental architectural design, the disaggregated analysis in Table 6 and the qualitative recovery of the network topology in Figure 2 allow us to dissect *why* this approach succeeds where others falter.

A primary finding is our model's profound adaptability to structural heterogeneity. Unlike methods that impose a single, global assumption about task relatedness, such as a shared low-rank subspace common in trace norm regularization approaches, our model learns a full, rich covariance matrix Σ, this allows it to simultaneously accommodate tightly-coupled task clusters (C1, C3), moderately related groups (C2), and near-independent tasks (C4), the successful performance on the "Outlier Tasks" is particularly revealing. Here, HB-MTL effectively learned to isolate these tasks, performing on par with the independent baseline and thus avoiding the classic MTL pitfall of negative transfer, where forcing irrelevant information sharing degrades performance.

Perhaps the most significant contribution demonstrated is the model's ability to effectively model and leverage negative correlations. Most MTL frameworks are built upon notions of "similarity" or "sharing," which implicitly assumes positive relatedness, the catastrophic failure of other MTL models on the "Anti-correlated Tasks" (C5) highlights this systemic weakness. Our model, by placing a flexible Inverse-Wishart prior on the covariance matrix, is unconstrained in this regard, it can infer negative off-diagonal entries in Σ if the data supports it, thereby correctly learning that an increase in traffic on a downtown arterial predicts a corresponding increase on an alternate route, this capability is a significant departure from standard MTL and is crucial for modeling complex systems where oppositional or competitive relationships exist.

Furthermore, the model's Bayesian nature offers critical advantages beyond predictive accuracy, the ability to visualize the posterior mean of the correlation matrix (Figure 2) transforms the model from a "black box" into an interpretable, knowledge-discovery tool. For a domain expert, such as a city planner, this inferred relationship map is an actionable output in itself. Equally important is the quantification of uncertainty. As shown in Table 8, the model's ability to distinguish between high confidence (for in-distribution data) and low confidence (for out-of-distribution data) is a hallmark of robust Bayesian inference, this epistemic uncertainty is vital for risk-sensitive applications, providing a clear signal to a user or an automated system about when a prediction can be trusted.

The primary limitation of our approach lies in its computational complexity, the Coordinate Ascent Variational Inference (CAVI) algorithm, while effective, can be computationally intensive for problems with a very large number of tasks (T) or high-

dimensional features (d), this presents a trade-off between model flexibility and scalability. Moreover, the choice of hyper-priors, while intended to be weakly informative, can still exert influence, especially in low-data regimes.

In summary, the discussion of our results validates the central thesis of this work: by moving away from fixed structural assumptions and embracing a fully adaptive Bayesian hierarchy, we can build MTL models that are more accurate, more interpretable, and safer for real-world deployment.

Conclusion

In this paper, we proposed and validated a Hierarchical Bayesian Model for Adaptive Multi-Task Learning (HB-MTL). Our approach formulates the MTL problem within a coherent probabilistic framework that learns, rather than assumes, the statistical structure shared among tasks. By inferring a full covariance matrix from the data, the model can capture complex, heterogeneous relationships, including the often-overlooked but critical phenomenon of negative correlation. Our comprehensive experimental evaluation on the "MetroSim" dataset demonstrated the clear advantages of this adaptivity, the model not only achieved a lower overall error than a range of competing methods but also excelled in navigating the nuanced substructures of the problem, correctly leveraging positive and negative relationships while avoiding the performance degradation of negative transfer. The contributions of this work are threefold: (1) a highly accurate predictive model for complex MTL problems; (2) an interpretable framework that provides a "map" of the inferred task relationships; and (3) a robust mechanism for quantifying predictive uncertainty, which is essential for reliable decision-making. Future work will focus on improving the scalability of the inference process, potentially through stochastic or distributed variational methods, and extending the framework to handle non-Gaussian likelihoods for classification and count-based tasks.

References

Baek J., Lesmes L., Lu Z. L. (2014). Bayesian adaptive estimation of the sensory memory decay function: The quick partial report method. Journal of Vision, 14 (10): 157, doi:10.1167/14.10.157.

Cavagnaro D. R., Myung J. I., Pitt M. A., Kujala J. V. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. Neural Computation, 22 (4), 887–905.

DiMattina C. (2015). Fast adaptive estimation of multi-dimensional psychometric functions. Journal of Vision, 15 (9): 15 1–20, doi:10.1167/15.9.5.

DiMattina C., Zhang K. (2008). How optimal stimuli for sensory neurons are constrained by network architecture. Neural Computation, 20 (3), 668–708.

DiMattina C., Zhang K. (2011). Active data collection for efficient estimation and comparison of nonlinear neural models. Neural Computation, 23 (9), 2242–2288.

Dorr M., Lesmes L. A., Lu Z. L., Bex P. J. (2013). Rapid and reliable assessment of the contrast sensitivity function on an iPad. Investigative Ophthalmology & Visual Science, 54 (12), 7266–7273.

Hou F., Huang C. B., Lesmes L., Feng L. X., Tao L., Zhou Y. F., Lu Z.-L. (2010). qCSF in clinical application: Efficient characterization and classification of contrast sensitivity functions in amblyopia. Investigative Ophthalmology & Visual Science, 51 (10), 5365–5377.

Hou F., Lesmes L., Bex P., Dorr M., Lu Z.-L. (2015). Using 10AFC to further improve the efficiency of quick CSF method. Journal of Vision, 15 (9): 15 1–18, doi:10.1167/15.9.2.

Hou F., Lesmes L., Kim W., Gu H., Pitt M., Myung J., Lu Z.-L. (2016). The usefulness of the quick CSF method: A large sample study. Manuscript submitted for publication.

Hou F., Lu Z.-L., Huang C. B. (2014). The external noise normalized gain profile of spatial vision. Journal of Vision, 14 (13): 15 1–14, doi:10.1167/14.13.9.

Huang C., Tao L., Zhou Y., Lu Z. L. (2007). Treated amblyopes remain deficient in spatial vision: A contrast sensitivity and external noise study. Vision Research, 47 (1), 22–34.

Kim W., Pitt M. A., Lu Z.-L., Steyvers M., Myung J. I. (2014). A hierarchical adaptive approach to optimal experimental design. Neural Computation, 26, 2463–2492.

Kleiner M., Brainard D., Pelli D., Ingling A., Murray R., Broussard C. (2007). What's new in Psychtoolbox-3. Perception, 36 (14), 1–16.

Kujala J. V., Lukka T. J. (2006). Bayesian adaptive estimation: The next dimension. Journal of Mathematical Psychology, 50 (4), 369–389.

Lee M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. Cognitive Science, 30 (3), 1–26.

Lesmes L. A., Jeon S. T., Lu Z. L., Dosher B. A. (2006). Bayesian adaptive estimation of threshold versus contrast external noise functions: The quick TvC method. Vision Research, 46 (19), 3160–3176.

Lesmes L. A., Lu Z. L., Baek J., Albright T. D. (2010). Bayesian adaptive estimation of the contrast sensitivity function: The quick CSF method. Journal of Vision, 10 (3): 15 1–21, doi:10.1167/10.3.17.

Lesmes L. A., Lu Z.-L., Baek J., Tran N., Dosher B. A., Albright T. D. (2015). Developing Bayesian adaptive methods for estimating sensitivity thresholds (d′) in Yes-No and forced-choice tasks. Frontiers in Psychology, 6, 1070, doi:10.3389/fpsyg.2015.01070.

Lu Z.-L., Dosher B. A. (2013). Visual psychophysics: From laboratory to theory. Cambridge, MA: MIT Press.

McAnany J. J., Alexander K. R. (2006). Contrast sensitivity for letter optotypes vs. gratings under conditions biased toward parvocellular and magnocellular pathways. Vision Research, 46 (10), 1574–1584.

Oshika T., Okamoto C., Samejima T., Tokunaga T., Miyata K. (2006). Contrast sensitivity function and ocular higher-order wavefront aberrations in normal human eyes. Ophthalmology, 113 (10), 1807–1812.

Rodriguez A., Dunson D. B., Gelfand A. E. (2008). The nested Dirichlet process. Journal of the American Statistical Association, 103, 1131–1154.

Teh Y. W., Jordan M. I., Beal M. J., Blei D. M. (2006). Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101, 1566–1581.

Wackerly D., Mendenhall W., Scheaffer R. (2007). Mathematical statistics with applications. Belmont, CA: Cengage Learning.

Wagenmakers E. J., Lee M., Lodewyckx T., Iverson G. J. (2008). Bayesian versus frequentist inference. In Hoijtink H., Klugkist I., Boelen P. A. (Eds.), Bayesian evaluation of informative hypotheses (pp. 181–207). New York: Springer.