arXiv:2305.00934v1 [stat.ML] 1 May 2023

预印本(2023) **05**，第 05 期，第

模型不确定性和参数不确定性下贝叶斯神经网络的变分推断

Aliaksandr Hubin *，Geir Storvik †

摘要。由于可扩展的贝叶斯近似推断技术的发展，贝叶斯神经网络（贝叶斯神经网络）近年来重新受到深度学习领域的广泛关注。采用贝叶斯方法具有若干优势：参数和预测不确定性能够轻松获得，从而促进严谨的统计分析。此外，还可以纳入先验知识。然而，迄今为止，尚无能够同时结合结构不确定性和参数不确定性的可扩展技术。本文将模型不确定性的概念作为贝叶斯神经网络结构学习的框架，从而在结构/模型与参数的联合空间中进行推断。此外，我们提出对一种可扩展的变分推断方法进行调整，通过对边际纳入概率进行重参数化，以纳入模型空间约束。在多个基准数据集上的实验结果表明，与竞争模型相比，我们获得了相当的准确率结果，但所采用的方法比普通贝叶斯神经网络更加稀疏。

关键词：贝叶斯神经网络；结构学习；模型选择；模型平均；近似贝叶斯推断；预测不确定性。

1 引言

近年来，频率学派深度学习方法已广受欢迎，并在从自然语言到图像分析的广泛现实应用中取得了巨大成功（Goodfellow等人，2016年）。这些算法通过迭代应用一系列非线性变换，旨在根据外层特征实现对响应变量的最优预测，从而能够灵活地对响应的复杂条件分布进行建模。每次变换都会产生另一隐藏层特征，这些特征也称为神经元。深度神经网络的架构/结构包括对层内非线性变换（激活函数）、层数（深度）、每层特征数（宽度）以及神经元之间连接（权重）的设定。在标准（频率学派）设置中，会使用某种优化方法（例如随机梯度下降）针对模型参数进行训练，以优化特定目标（例如最小化均方根误差或负对数似然）。深度学习方法往往优于传统统计模型，即使后者经过精心设计并融入了专家知识（Refenes等人，1994；Razi和Athappilly，2005；Adya和Collopy，1998；Sargent，2001；Kanter和Veeramachaneni，2015年）。然而，通常需要使用庞大的数据集，才能构建具有泛化能力的神经网络并避免过拟合问题。尽管人们已为频率学派深度学习方法开发了若干正则化技术（L_1 和 L_2 权重惩罚、丢弃法、批归一化等），以避免对训练数据集过拟合，但这些方法是否奏效并不显而易见。研究表明，可以通过将某些权重置零或移除某些节点，以非结构化方式对网络进行剪枝。

*奥斯陆大学数学系，NMBU，NR，HiOF。aliaksah@math.uio.no
†奥斯陆大学数学系，NR。geirs@math.uio.no

© 2023 作者

作为频率学派深度学习方法的一种替代方案，贝叶斯神经网络代表了一类非常灵活的模型，具有较强的抗过拟合能力（Neklyudov 等人，2018年）。然而，这类模型通常仍然存在严重的参数过度化问题。目前已有多种隐式方法，通过先验分布对权重进行收缩，从而实现贝叶斯神经网络的稀疏化（Jylänki 等人，2014年；Blundell 等人，2015年；Molchanov 等人，2017年；Ghosh 等人，2018年；Neklyudov 等人，2017年）。例如，Blundell 等人（2015年）提出了两个高斯密度的混合，随后采用完全因子化的平均场变分近似。Ghosh 等人（2019年）和 Louizos 等人（2017年）分别通过为权重引入马蹄先验（Carvalho 等人，2009年），对这一方法进行了推广，从而实现了更强的收缩效果，并自动确定了 Blundell 等人（2015年）所需的混合成分方差。某些算法过程也可以被视为对应于特定的贝叶斯先验。例如，Molchanov 等人（2017年）证明，高斯 dropout 等价于权重参数服从对数均匀先验的贝叶斯神经网络。

本文中，我们考虑一种形式化的贝叶斯方法，联合考虑结构不确定性和参数不确定性，将其作为针对线性回归模型中的贝叶斯模型选择而开发的方法在贝叶斯神经网络中的推广。该方法通过引入与给定架构中特定权重的纳入或排除相对应的潜在二元变量来实现。具体做法是引入尖峰与 slab 先验。针对贝叶斯神经网络情形的此类先验由Polson 和 Rockova(2018) 以及 Hubin (2018) 提出。本文早期版本中提出了这类情形下的推断计算过程（Hubin 和 Storvik, 2019），随后 Bai 等人(2020) 未作任何修改地采用了该过程。 Bai 等人(2020) 还给出了关于先验纳入概率选择的渐近理论结果，但该结果仅在参数数量趋于无穷且 slab 成分的先验方差固定时成立。本文进一步引入先验纳入概率和 slab 成分方差的超先验，使其成为随机量；同时，我们还允许采用基于纳入指示变量的多元高斯结构的、更灵活的变分近似。此外，我们考虑了若干用于完全贝叶斯模型平均的替代预测方法，包括基于后验均值的模型和中位概率模型，并开展了一项全面的实验研究，将所提出的方法与若干竞争算法及多个数据集进行比较。

在模型空间中采用贝叶斯形式化，可以将贝叶斯推断的整套机制应用于模型与参数联合设定，包括 贝叶斯模型平均（ BMA ）（跨越所有模型）或 贝叶斯模型选择 （ BMS ），依据某种模型选择准则选择一个 “ 最佳 ” 模型（ Claeskens 等人，2008年）。本文研究 BMA，以及 中位概率 模型（Barbieri 等人，2004，2021年）和基于后验均值模型的 BNNs 推断。实验中还讨论了 BMS（尤其是中位概率模型）的稀疏化特性。最后，遵循 Hubin （2018年），我们将所得的边际纳入概率 与二元 *dropout* 率联系起来，从而为后者提供恰当的概率推理。推断算法基于可扩展的随机变分推断。

所提出的方法与已广受欢迎的二元 dropout 有相似之处（Srivastava 等人，2014年）。然而，标准二元 dropout 可以被视为高斯过程模型的贝叶斯近似，其中仅考虑参数估计（Gal，2016年），而我们的方法明确地建模 结构不确定性。从这个意义上说，它与 Concrete dropout（Gal 等人，2017年）密切相关。然而，Gal 等人（2017年）提出的模型不允许进行 BMS：由于假设整个层 具有相同的 dropout 概率，中位概率模型要么选择全部权重，要么一个也不选择。此外，在Gal 等人（2017年）中应用的变分近似过程尚未在模型不确定性的语境下得到研究。

同时，必须明确指出，除广义线性模型中的一种特殊模型选择情形外，我们的方法目前在大多数情况下并不旨在对贝叶斯神经网络进行可解释推断。原则上，如果允许所有层都直接连接到响应变量，那么可解释模型或许是可行的。然而，本文不讨论这类情形。

本文其余部分安排如下：第2 节介绍一些背景知识并讨论相关研究。第3节从数学上定义贝叶斯神经网络类及其对应的模型空间。第4节介绍如何利用边际纳入概率的重参数化算法训练所提出的模型类。第4.3节讨论若干预测推断方法。第5节将所提出的方法应用于两个经典基准数据集 MNIST、FMNIST（用于图像分类）以及 PHONEME（用于声音分类）。我们还将结果与一些现有的贝叶斯神经网络推断方法进行比较。最后，第6 节总结全文并提出进一步研究的建议。本文的补充材料中还提供了其他结果。

2 背景与相关工作

贝叶斯神经网络（BNN）由 Neal（1992年）、MacKay（1995年）和 Bishop（1997年）于几十年前首次提出。贝叶斯神经网络利用严谨的贝叶斯方法，能够妥善处理参数和预测不确定性，原则上还可以纳入先验知识。在许多情况下，这能带来更稳健、过拟合程度更低的解决方案。然而，这需要付出极高的计算成本。直到最近，由于标准马尔可夫链蒙特卡罗方法的局限性（它是当时主要使用的数值计算方法），对贝叶斯神经网络进行推断仍无法扩展到大型且高维的数据集。已有研究提出了多种基于马尔可夫链蒙特卡罗子采样技术的方法，包括近似方法（Bardenet 等，2014年、2017年；Korattikara 等，2014年；Quiroz 等，2014年；Welling 和 Teh，2011年）和精确方法（Quiroz 等，2016年；Maclaurin 和 Adams，2014年；Liu 等，2015年；Welling 和 Teh，2011年），但在超高维场景中，这些方法都无法高效探索参数空间。

MCMC技术的一种替代方法是通过变分贝叶斯（也称为变分推断）进行近似贝叶斯推断（Jordan等人，1999年）。由于变分方法具有快速收敛的特性，在高维问题中，变分推断算法通常比MCMC算法快几个数量级（Ahmed等人，2012年）。变分推断在潜变量模型中有多种应用，例如混合模型（Humphreys和Titterington，2000年）、隐马尔可夫模型（MacKay，1997年）以及一般的图模型（Attias，2000年）。Graves (2011年)提出了将可扩展变分推断应用于贝叶斯神经网络的方法。通过纳入多种方差缩减技术，这一方法得到了进一步改进，相关讨论见Gal (2016年)。

正如引言中所述，研究表明，神经网络中的大多数权重参数都可以从模型中剪除，而不会显著降低预测准确率。然而，剪枝通常是通过启发式阈值化隐式完成的，即删除权重。不过，学习模型中应包含哪些参数，也可以被视为结构学习或模型选择问题。

正如Claeskens 和 Hjort (2008)、Steel (2020) 或 Hansen 和 Yu (2001)（以及其他文献）所讨论的那样，统计学中的模型选择和模型平均通常假设存在一个离散且可数（有限或无限）的模型集合，并且这些模型位于相应的模型空间中。模型空间中的模型可能在所采用的似然、所涉及的链接函数，或线性或非线性预测器中包含哪些参数方面有所不同。模型选择的目的是从模型空间中选出一个单一模型（在某种意义上最优的模型）。在某些应用中，这种选择可以得到更易于解释的模型（例如线性回归中的变量选择，Kuo 和 Mallick，1998），而在其他应用中则只是为了某种目的（例如预测）选择最佳模型（Geisser 和 Eddy，1979）；有时两者会重合（Hubin 等人，2021），但并非总是如此（Breiman，2001）。模型选择中的稀疏性是否重要取决于具体情境，但通常希望在某种意义上做到简约。与此同时，模型选择往往会引发不确定性处理方面的问题，导致参数的置信区间/可信区间过窄（Heinze 等人，2018），因而也常会导致预测出现类似问题，即过拟合。模型平均（如果模型不确定性

得到妥善处理）可以解决这些问题（Bornkamp 等人，2017年），并具有其他优势（Steel，2020）。此外，考虑模型不确定性的模型选择，例如使用中位概率模型，能够更有效地抵御过拟合（Ghosh，2015年）。

已有大量研究表明，通过引入与不同离散模型配置相对应的潜变量，模型选择/平均在参数选择方面既高效又准确。在贝叶斯语境下，随后可利用后验分布选择最佳稀疏配置，并明确处理联合模型与参数不确定性（George 和 McCulloch，1993；Clyde 等人，2011年；Frommlet 等人，2012年；Hubin 和 Storvik，2018；Hubin 等人，2020，2021）。尖峰与板块先验（Mitchell 和 Beauchamp，1988）通常用于此类情形。这些方法均展现出良好的稀疏模型预测性能，并能够恢复有意义的复杂非线性。不过，它们基于对马尔可夫链蒙特卡罗（MCMC）的改造，难以扩展至大规模高维数据样本。Louizos 等人（2017年）还指出，在贝叶斯神经网络中对模型配置进行显式离散化十分复杂，因为这会导致参数总数呈指数级增长，从而使高维问题的推断变得不可行。Polson 和 Rockova （2018）从理论角度研究了尖峰与板块方法在贝叶斯神经网络中的应用。

Logsdon 等人 （2010年）；Carbonetto 等人 （2012年）提出了一种完全因子化变分分布，能够在全基因组关联研究（GWAS）中典型的、具有超高数量潜在协变量的线性模型中，高效且精确地“线性化”贝叶斯模型选择的计算负担。在其博士论文的讨论中，Hubin （2018年）提出结合Logsdon 等人 （2010年）；Carbonetto 等人 （2012年）和 Graves （2011年）的方法，以便在深度贝叶斯回归模型中，对模型与参数的联合空间进行可扩展的贝叶斯近似推断。本文将进一步发展这一思路。

3 模型

神经网络模型通过概率函数映射将（可能是多维的）观测值 $\boldsymbol{y}_i \in \mathcal{R}^r$ 与解释变量 $\boldsymbol{x}_i \in \mathcal{R}^p$ 联系起来，该映射具有均值参数向量 $\boldsymbol{\mu}_i = \boldsymbol{\mu}_i(\boldsymbol{x}_i) \in \mathcal{R}^r$:

$$\boldsymbol{y}_i \sim \mathfrak{f}\left(\boldsymbol{\mu}_i(\boldsymbol{x}_i), \phi\right), \quad i \in \{1, ..., n\}, \tag{3.1}$$

其中， $\mathfrak{f}$ 是某种观测分布，通常属于指数族，而 ϕ 是离散参数。为构造均值参数向量 $\boldsymbol{\mu}_i$，需要通过半仿射变换，依次构建由隐藏层组成的模块:

$$z_{ij}^{(l+1)} = g_j^{(l)}\left(\beta_{0j}^{(l)} + \sum_{k=1}^{p^{(l)}} \beta_{kj}^{(l)} z_{ik}^{(l)}\right), l = 1, ..., L-1, j = 1, ..., p^{(l+1)}, \tag{3.2}$$

其中 $\mu_{ij} = z_{ij}^{(L)}$ 。这里， L 为层数， $p^{(l)}$ 为相应层中的节点数，而 $g_j^{(l)}$ 是单变量函数（下文称为激活函数）。此外，$\beta_{kj}^{(l)} \in \mathcal{R}$， $k > 0$ 是第 l 层输入 $z_{ik}^{(l)}$ 的权重（斜率系数）（注意， $z_{ik}^{(1)} = x_{ik}$ 和 $p^{(1)} = p$）。对于 $k = 0$，我们得到截距/偏置项。最后，我们引入潜在二元指示变量 $\gamma_{kj}^{(l)} \in \{0, 1\}$ ，用于开启或关闭相应的权重，使得当 $\gamma_{kj}^{(l)} = 0$时，$\beta_{kj}^{(l)} = 0$ 。

在我们的记号中，我们明确区分由向量$\boldsymbol{\gamma} = \{\gamma_{kj}^{(l)}, j = 1, .., p^{(l+1)}, k = 0, ..., p^{(l)}, l = 1, ..., L-1\}$ 定义的离散结构/模型配置（下文称为模型），它们构成模型空间 Γ，以及

在这些配置条件下的模型参数 $\boldsymbol{\theta}|\gamma = \{\boldsymbol{\beta}, \phi|\boldsymbol{\gamma}\}$，其中仅纳入那些 $\beta_{kj}^{(l)}$ 满足 $\gamma_{kj}^{(l)} = 1$ 的参数。这种方法在统计科学文献中相当常见，用于明确说明给定模型类别中的模型不确定性，例如见Clyde 等人(2011年)；Frommlet 等人(2012年)；Hubin 等人(2021)。

通过为每个模型指定模型先验分布 $p(\boldsymbol{\gamma})$ 和参数先验分布 $p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \phi)$，即可完成贝叶斯方法的设定。如果结果的分布中包含离散参数，还需要为其定义先验分布 $p(\phi|\boldsymbol{\gamma})$。对于 $p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \phi)$，可以考虑多种先验分布，包括高斯混合先验（Blundell 等人，2015年）、马蹄先验（Ghosh 等人，2019年；Louizos 等人，2017年）或 g 先验混合（Li 和 Clyde，2018年），这些先验可以对权重参数施加额外惩罚。我们首先沿用早期预印本（Hubin 和 Storvik，2019年）中的做法，同时参考 Hubin（2018年）以及 Polson 和 Rockova（2018年）更早提出的思想，考虑将独立高斯尖峰与 slab 权重先验与独立伯努利先验相结合，使潜在纳入指示变量取值为 1。这种先验分布选择对应于权重的边际尖峰与 slab 先验（Clyde 等人，2011年）：

$$p(\beta_{kj}^{(l)}|\sigma_{\beta,l}^2, \gamma_{kj}^{(l)}) = \gamma_{kj}^{(l)}\mathcal{N}(0, \sigma_{\beta,l}^2) + (1 - \gamma_{kj}^{(l)})\delta_0(\beta_{kj}^{(l)}), \tag{3.3a}$$

$$p(\gamma_{kj}^{(l)}) = \text{Bernoulli}(\psi^{(l)}), \tag{3.3b}$$

此处，$\delta_0(\cdot)$表示零点处的狄拉克质量或“尖峰”，$\sigma_{\beta,\,l}^2$ 是 $\beta_{kj}^{(l)}$ 的先验方差，而 $\psi^{(l)} \in (0,\ 1)$是将权重 $\beta_{kj}^{(l)}$ 纳入模型的先验概率。

为了自动推断先验方差以及将权重纳入模型的先验概率，我们假设$\sigma_{\beta,\,l}^2$ 服从具有超参数 $a_\beta^{(l)}$，$b_\beta^{(l)}$ 的标准逆伽马超先验，而 $\psi^{(l)}$ 服从 Beta($a_\psi^{(l)}$，$b_\psi^{(l)}$)先验：

$$p(\beta_{kj}^{(l)}|\sigma_{\beta,l}^2, \gamma_{kj}^{(l)}) = \gamma_{kj}^{(l)}\mathcal{N}(0, \sigma_{\beta,l}^2) + (1 - \gamma_{kj}^{(l)})\delta_0(\beta_{kj}^{(l)}), \tag{3.4a}$$

$$p(\sigma_{\beta,l}^2) = \text{Inv-Gamma}(a_\beta^{(l)}, b_\beta^{(l)}), \tag{3.4b}$$

$$p(\gamma_{kj}^{(l)}) = \text{Bernoulli}(\psi^{(l)}), \tag{3.4c}$$

$$p(\psi^{(l)}) = \text{Beta}(a_\psi^{(l)}, b_\psi^{(l)}). \tag{3.4d}$$

此处，$\delta_0(\cdot)$ 表示零点处的狄拉克质量或“尖峰”，$\sigma_{\beta,\,l}^2$ 是 $\beta_{kj}^{(l)}$ 的先验方差，服从具有超参数 $a_\beta^{(l)}$，$b_\beta^{(l)}$ 的标准逆伽马超先验，而 $\psi^{(l)} \in [0, 1]$ 是将权重 $\beta_{kj}^{(l)}$ 纳入模型的先验概率。此外，假设 $\psi^{(l)}$ 服从 Beta 分布($a_\psi^{(l)}$，$b_\psi^{(l)}$)。存在离散参数时，需要额外的先验，参见Dey 等人(2000年。我们将此模型称为潜在二元贝叶斯神经网络（LBBNN）模型。

4 贝叶斯推断

在模型和参数均存在不确定性的情况下，推断的主要目标是根据数据 $\mathcal{D}$推断某个感兴趣的参数 Δ 的后验边缘分布（例如，在给定新的协变量 $\boldsymbol{x}^*$的条件下，新观测 y^* 的分布）：

$$p(\Delta|\mathcal{D}) = \sum_{\boldsymbol{\gamma}\in\Gamma} \int_{\boldsymbol{\theta}\in\Theta_\gamma} p(\Delta|\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathcal{D}) p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathcal{D}) \mathrm{d}\boldsymbol{\theta}, \tag{4.1}$$

其中 Θ_γ 是通过 γ定义的参数空间。处理复杂后验分布的标准方法是应用蒙特卡罗方法，即从 $p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathcal{D})$进行仿真。对于由 (3.1)-(3.4定义且包含许多隐变量的模型，此类仿真会变得棘手。

背后的想法　和变分推断（Graves，2011年；Blei等人，2017年）是应用该近似

$$\tilde{p}(\Delta|\mathcal{D}) = \sum_{\boldsymbol{\gamma}\in\Gamma} \int_{\boldsymbol{\theta}\in\Theta_\gamma} p(\Delta|\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathcal{D}) q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) \mathrm{d}\boldsymbol{\theta}. \tag{4.2}$$

对于某些合适的（参数化）分布 $q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma})$，通过对参数 $\boldsymbol{\eta}$ 进行适当选择，使其能够良好地近似后验分布，并且易于从中采样。 $\boldsymbol{\eta}$ 的具体形式通常通过最小化从变分族分布到后验分布的 Kullback-Leibler散度来确定：

$$\mathrm{KL}(q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma})||p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathcal{D})) = \sum_{\boldsymbol{\gamma}\in\Gamma} \int_{\Theta_\gamma} q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) \log \tfrac{q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathcal{D})} \mathrm{d}\boldsymbol{\theta}, \tag{4.3}$$

关于变分参数 $\boldsymbol{\eta}$。与标准变分推断方法相比，此处的设定扩展为包含离散模型标识符 $\boldsymbol{\gamma}$。对于最优选择$\hat{}\boldsymbol{\eta}$ 的 $\boldsymbol{\eta}$，通过对（ Δ ）进行蒙特卡罗估计来执行推断4.2，并代入$\hat{}\boldsymbol{\eta}$ 以替代 $\boldsymbol{\eta}$。因此，主要挑战在于选择合适的变分族，以及用于最小化（4.3）的计算过程。需要注意的是，尽管这一最小化仍然面临计算挑战，但通常会比直接从真实后验分布中获取样本容易得多。只要选择的变分分布$q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ 易于采样，最终的蒙特卡罗估计就会很简单。

与变分推断的标准设置一样，散度（4.3）的最小化等价于证据下界（ELBO）的最大化

$$\mathcal{L}_{VI}(\boldsymbol{\eta}) = \sum_{\boldsymbol{\gamma}\in\Gamma} \int_{\Theta_\gamma} q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) \log p(\mathcal{D}|\boldsymbol{\theta}, \boldsymbol{\gamma}) \mathrm{d}\boldsymbol{\theta} - \mathrm{KL}\left(q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma})||p(\boldsymbol{\theta}, \boldsymbol{\gamma})\right) \tag{4.4}$$

通过以下等式

$$\mathcal{L}_{VI}(\boldsymbol{\eta}) = p(\mathcal{D}) - \mathrm{KL}\left(q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma})||p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathcal{D})\right),$$

这也表明， $\mathcal{L}_{VI}(\boldsymbol{\eta})$ 是边际似然 $p(\mathcal{D})$ 的下界。

4.1 变分分布

我们将考虑一种先前针对线性回归（Logsdon 等，2010年；Carbonetto 等，2012年）提出的变分族，并将其扩展到 LBBNN 设置。假设

$$q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = q_{\boldsymbol{\eta}_0}(\phi) \prod_{l=1}^{L-1} \prod_{j=1}^{p^{(l+1)}} \prod_{k=0}^{p^{(l)}} q_{\kappa_{kj}, \tau_{kj}}(\beta_{kj}^{(l)}|\gamma_{kj}^{(l)}) q_{\alpha_{kj}^{(l)}}(\gamma_{kj}^{(l)}), \tag{4.5}$$

其中，$q_{\boldsymbol{\eta}_0}(\phi)$ 是一个适用于离散参数的分布，

$$q_{\kappa_{kj}^{(l)}, \tau_{kj}^{(l)}}\left(\beta_{kj}^{(l)}|\gamma_{kj}^{(l)}\right) = \gamma_{kj}^{(l)} \mathcal{N}(\kappa_{kj}^{(l)}, {\tau^2}_{kj}^{(l)}) + (1 - \gamma_{kj}^{(l)}) \delta_0(\beta_{kj}^{(l)}), \tag{4.6}$$

and

$$q_{\alpha_{kj}^{(l)}}(\gamma_{kj}^{(l)}) = \text{Bernoulli}(\alpha_{kj}^{(l)}). \tag{4.7}$$

以概率 $\alpha_{kj}^{(l)} \in [0,1]$，权重 $\beta_{kj}^{(l)}$ 的参数后验分布将通过具有某个均值和方差的正态分布（“厚片”）进行近似；否则，该权重将被置为零。因此，$\alpha_{kj}^{(l)}$ 将近似表示权重 $\beta_{kj}^{(l)}$ 的边际后验纳入概率。这里，$\boldsymbol{\eta} = \{\boldsymbol{\eta}_0, (\kappa_{kj}^{(l)}, \tau^{2^{(l)}}_{kj}, \alpha_{kj}^{(l)})\ l = 1, ... L-1, \ k = 1, ... p^{(l+1)}, \ j = 1, ... p^{(l)}\}$。

通过丢弃法，贝叶斯神经网络中也采用了类似的变分分布（Srivastava 等人，2014年）。然而，对于丢弃法，最终的网络是稠密的，但训练时采用的是对稀疏网络进行蒙特卡罗平均的方式。在我们的方法中，目标分布有所不同，因为其中包含二元变量 $\{\gamma_{kj}^{(l)}\}$ 作为模型的一部分。因此，我们的边际纳入概率可视为丢弃率的一种特殊情形，并且对结构模型不确定性具有恰当的概率解释。

变分分布（4.5）-（4.7）对应于常用的平均场近似，可视为一种相当粗糙的近似，它完全忽略了模型结构或参数之间的所有后验依赖。因此，由此得出的结论可能具有误导性或不准确，因为某个权重的后验概率可能会受到其他权重是否纳入的显著影响。这种依赖结构可以通过 γ或 β（或两者）引入变分近似中。这里，我们只考虑纳入变量之间的依赖结构。我们仍假设不同层之间相互独立，但在层内，通过将 $\boldsymbol{\alpha}^l = \{\alpha_{kj}^{(l)}\}$ 定义为一个随机向量，并令其在 logit 尺度上服从多元正态分布，从而引入依赖结构：

$$\text{logit}(\boldsymbol{\alpha}^{(l)}) \sim \text{MVN}(\boldsymbol{\xi}^{(l)}, \boldsymbol{\Sigma}^{(l)}). \tag{4.8}$$

此处既可以使用完整协方差矩阵 $\boldsymbol{\Sigma}^{(l)}$，也可以对协方差进行低秩参数化。对于后一种情况，$\boldsymbol{\Sigma}^{(l)} = \boldsymbol{F}^{(l)}\boldsymbol{F}^{(l)^T} + \boldsymbol{D}^{(l)}$ 其中，$\boldsymbol{F}^{(l)}$ 是协方差矩阵低秩形式中的因子部分，而 $D^{(l)}$ 是协方差矩阵低秩形式中的对角部分。这大幅减少了参数数量，并可高效计算行列式和逆矩阵。一个特别有趣的情况是，当 $\boldsymbol{F}^{(l)}$ 的秩为零时，各分量之间仍保持独立，但 $\alpha_{kj}^{(l)}$ 的变异性会受到一定程度的惩罚。在参数化形式（4.6）-（4.8）下，参数 $\{\boldsymbol{\xi}^{(l)}, l = 1, ..., L-1\}$ 和 $\{\boldsymbol{\Sigma}^{(l)}, l = 1, ..., L-1\}$ 会被加入参数向量 $\boldsymbol{\eta}$中。随后，也会对这些参数应用重参数化技巧，使用 PyTorch probabilities 中为多元正态分布和 LFMVN 提供的默认表示，这些表示开箱即用。PyTorch 概率分布。

4.2 随机梯度优化

为简单起见，这里假设不存在离散参数，因此目标分布为 $p(\boldsymbol{\beta}, \ \boldsymbol{\gamma}|\mathcal{D})$。我们可以将证据下界（4.4）改写为

$$\mathcal{L}_{VI}(\boldsymbol{\eta}) = \sum_{\boldsymbol{\gamma}\in\Gamma} \int_{\beta_\gamma} q_{\boldsymbol{\eta}}(\boldsymbol{\beta}, \boldsymbol{\gamma})[\log p(\mathcal{D}|\boldsymbol{\beta}, \boldsymbol{\gamma}) - \log \tfrac{q_{\boldsymbol{\eta}}(\boldsymbol{\beta}, \boldsymbol{\gamma})}{p(\boldsymbol{\beta}, \boldsymbol{\gamma})}]\mathrm{d}\boldsymbol{\beta}. \tag{4.9}$$

由于当 Γ 和 $\mathcal{D}$ 较大时，计算梯度的计算成本极高，因此使用蒙特卡罗估计来获得无偏梯度估计的随机梯度方法，已成为

此类情况下变分推断的标准方法。重参数化技巧和小批量处理（Kingma 等人，2015年；Blundell 等人，2015年）也得到了进一步应用。

我们所讨论情形中的另一个难点是 γ的离散性质。遵循Gal 等人(2017年) 的做法，我们用Concrete 分布来放宽伯努利分布（*4.7*）：

$$\tilde{\gamma} = \gamma_{tr}(\nu, \delta; \alpha) = \text{sigmoid}((\text{logit}(\alpha) - \text{logit}(\nu))/\delta), \quad \nu \sim \text{Unif}[0, 1], \tag{4.10}$$

其中， δ 是一个调节参数，选取较小的值。在趋于零的极限下， $\tilde{\gamma}$ 会退化为伯努利(α)变量。结合对 β进行的重参数化，

$$\beta = \beta_{tr}(\varepsilon; \kappa, \tau) = \kappa + \tau\varepsilon, \quad \varepsilon \sim N(0, 1) \tag{4.11}$$

我们定义如下对证据下界的近似：

$$\begin{aligned}\mathcal{L}_{VI}^{\delta}(\boldsymbol{\eta}) := \int_{\boldsymbol{\nu}} \int_{\boldsymbol{\varepsilon}} q_{\boldsymbol{\nu},\boldsymbol{\varepsilon}}(\boldsymbol{\nu}, \boldsymbol{\varepsilon}) [\log p(\mathcal{D}|\beta_{tr}(\boldsymbol{\varepsilon}, \boldsymbol{\kappa}, \boldsymbol{\tau}), \gamma_{tr}(\boldsymbol{\nu}, \boldsymbol{\alpha}, \delta)) - \\ \log \tfrac{q_{\boldsymbol{\eta}}(\beta_{tr}(\boldsymbol{\varepsilon},\boldsymbol{\kappa},\boldsymbol{\tau}),\gamma_{tr}(\boldsymbol{\nu},\boldsymbol{\alpha},\delta))}{p(\beta_{tr}(\boldsymbol{\varepsilon},\boldsymbol{\kappa},\boldsymbol{\tau}),\gamma_{tr}(\boldsymbol{\nu},\boldsymbol{\alpha},\delta))}] \mathrm{d}\boldsymbol{\varepsilon}\mathrm{d}\boldsymbol{\nu}\end{aligned} \tag{4.12}$$

其中，对向量进行的变换均按元素逐一执行。此外，由于 $\mathrm{d}\boldsymbol{\varepsilon}\mathrm{d}\boldsymbol{\nu}$ 不依赖于 $\boldsymbol{\eta}$，因此在计算 $\mathcal{L}_{VI}^{\delta}(\boldsymbol{\eta})$的梯度时，我们可以交换积分与微分的顺序：

$$\begin{aligned}\nabla_{\boldsymbol{\eta}} \mathcal{L}_{VI}^{\delta}(\boldsymbol{\eta}) = \int_{\boldsymbol{\nu}} \int_{\boldsymbol{\varepsilon}} q_{\boldsymbol{\nu},\boldsymbol{\varepsilon}}(\boldsymbol{\nu}, \boldsymbol{\varepsilon}) \nabla_{\boldsymbol{\eta}} [\log p(\mathcal{D}|\beta_{tr}(\boldsymbol{\varepsilon}, \boldsymbol{\kappa}, \boldsymbol{\tau}), \gamma_{tr}(\boldsymbol{\nu}, \boldsymbol{\alpha}, \delta) - \\ \log \tfrac{q_{\boldsymbol{\eta}}(\beta_{tr}(\boldsymbol{\varepsilon},\boldsymbol{\kappa},\boldsymbol{\tau}),\gamma_{tr}(\boldsymbol{\nu},\boldsymbol{\alpha},\delta))}{p(\beta_{tr}(\boldsymbol{\varepsilon},\boldsymbol{\kappa},\boldsymbol{\tau}),\gamma_{tr}(\boldsymbol{\nu},\boldsymbol{\alpha},\delta))}] \mathrm{d}\boldsymbol{\varepsilon}\mathrm{d}\boldsymbol{\nu}.\end{aligned} \tag{4.13}$$

的$\boldsymbol{\eta}$（的无偏估计量由命题1给出。

命题 1. 假设对于所有 $m = 1, ..., M$ $\left(\boldsymbol{\nu}^{(m)}, \boldsymbol{\eta}^{(m)}\right) \sim q_{\boldsymbol{\nu},\boldsymbol{\varepsilon}}(\boldsymbol{\nu}, \boldsymbol{\eta})$，且 S 是一个从索引集合 $\{1, ..., n\}$ 中随机抽取的大小为 N 的子集。此外，假设观测值在条件上相互独立。那么，对于任意 $\delta > 0$， $\mathcal{L}_{VI}^{\delta}(\boldsymbol{\eta})$ 的梯度的无偏估计量为

$$\begin{aligned}\widetilde{\nabla}_{\boldsymbol{\eta}} \mathcal{L}_{VI}^{\delta}(\boldsymbol{\eta}) = \frac{1}{M} \sum_{m=1}^{M} \Big[\frac{n}{N} \sum_{i \in S} \nabla_{\boldsymbol{\eta}} \log p(\boldsymbol{y_i}|\boldsymbol{x_i}, \beta_{tr}(\boldsymbol{\varepsilon}^{(m)}, \boldsymbol{\kappa}, \boldsymbol{\tau}), \gamma_{tr}(\boldsymbol{\nu}^{(m)}, \boldsymbol{\alpha}, \delta)) - \\ \nabla_{\boldsymbol{\eta}} \log \tfrac{q_{\boldsymbol{\eta}}(\beta_{tr}(\boldsymbol{\varepsilon}^{(m)},\boldsymbol{\kappa},\boldsymbol{\tau}),\gamma_{tr}(\boldsymbol{\nu}^{(m)},\boldsymbol{\alpha},\delta))}{p(\beta_{tr}(\boldsymbol{\varepsilon}^{(m)},\boldsymbol{\kappa},\boldsymbol{\tau}),\gamma_{tr}(\boldsymbol{\nu}^{(m)},\boldsymbol{\alpha},\delta))} \Big].\end{aligned} \tag{4.14}$$

证明. 由 (4.13) 可得

$$\frac{1}{M} \sum_{m=1}^{M} \nabla_{\boldsymbol{\eta}} [\log p(\mathcal{D}|\beta_{tr}(\boldsymbol{\varepsilon}^{(m)}, \boldsymbol{\kappa}, \boldsymbol{\tau}), \gamma_{tr}(\boldsymbol{\nu}^{(m)}, \boldsymbol{\alpha}, \delta) - \log \tfrac{q_{\boldsymbol{\eta}}(\beta_{tr}(\boldsymbol{\varepsilon}^{(m)},,\boldsymbol{\kappa},\boldsymbol{\tau}),\gamma_{tr}(\boldsymbol{\nu}^{(m)},\boldsymbol{\alpha},\delta))}{p(\beta_{tr}(\boldsymbol{\varepsilon}^{(m)},\boldsymbol{\kappa},\boldsymbol{\tau}),\gamma_{tr}(\boldsymbol{\nu},^{(m)},\boldsymbol{\alpha},\delta))}]$$

是梯度的无偏估计。此外，由于我们假设观测值在条件上相互独立，因此有

$$\nabla_{\boldsymbol{\eta}} \log p(\mathcal{D}|\beta_{tr}(\boldsymbol{\varepsilon}, \boldsymbol{\kappa}, \boldsymbol{\tau}), \gamma_{tr}(\boldsymbol{\nu}, \boldsymbol{\alpha}, \delta)) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\eta}} \log p(\boldsymbol{y}_i|\boldsymbol{x}_i; \beta_{tr}(\boldsymbol{\varepsilon}, \boldsymbol{\kappa}, \boldsymbol{\tau}), \gamma_{tr}(\boldsymbol{\nu}, \boldsymbol{\alpha}, \delta)),$$

对于该式，一个 可以通过随机子集构造无偏估计量，从而得到该结果。 □

算法 1 双重随机变分推断步骤

从 N中均匀抽取索引，取值范围为 $\{1,...,n\}$，定义 S;
对于 m **中的** $\{1,...,M\}$，**执行**
  对于 $(k,j,l) \in \mathcal{B}$，**执行**
    从 $\nu_{kl}^{(l)} \sim$ 中采样 Unif$[0,1]$ 和 $\varepsilon_{kj}^{(l)} \sim N(0,1)$
  结束循环
结束循环
计算 $\widetilde{\nabla}_{\boldsymbol{\eta}} \mathcal{L}_{VI}^{\delta}(\boldsymbol{\eta})$ 根据（4.14）
更新 $\boldsymbol{\eta} \leftarrow \boldsymbol{\eta} + \boldsymbol{A} \widetilde{\nabla}_{\boldsymbol{\eta}} \mathcal{L}_{VI}^{\delta}(\boldsymbol{\eta})$

算法1 描述了双重随机变分推断方法的一次迭代。在均值场假设下（为简化起见），该方法对参数进行更新。集合 $\mathcal{B}$ 是网络中所有组合的集合j，k，l。学习率矩阵 $\boldsymbol{A}$ 始终为对角矩阵，因此可以为涉及的参数设置不同的步长。按照 Blundell 等人(2015年) 的做法，通过重参数化 $\tau_{kj}^{(l)}$ log($\tau_{kj}^{(l)} =$ exp$(1+)$) 其中$\rho_{kj}^{(l)}$。通常，更新会在完整的训练轮次中进行，此时观测值会被划分为 n/N 个子集，并依次对所有子集进行更新。

对于依赖结构（4.8），$\boldsymbol{\alpha}$ 改为进行抽样，而 $\boldsymbol{\xi}^{(l)}$ 及 $\boldsymbol{\Sigma}^{(l)}$ 的各个分量则代入 $\boldsymbol{\eta}$。对 $\alpha_{kj}^{(l)}$ 的约束通过重参数化 $\alpha_{kj}^{(l)} = (1+\exp(-\omega_{kj}^{(l)}))^{-1}$ 来纳入，其中 $\omega_{kj}^{(l)} \in \mathcal{R}$

请注意，在所提出的算法中，边际纳入概率的偏导数，以及权重的均值和标准差项，均可通过神经网络上的常规反向传播算法计算。该算法假设色散参数 ϕ 已知，但也很容易推广到同时学习 ϕ。

4.3 预测

获得变分近似分布的参数 $\boldsymbol{\eta}$ 的估计值 $\widehat{\boldsymbol{\eta}}$ 后，我们回到 $\boldsymbol{\gamma}$ 的原始离散模型（设定 $\delta = 0$）。接下来，可以采用多种方式进行预测推断。下面列出这些方法。

完全贝叶斯模型平均 在这种情况下，定义

$$\hat{p}(\Delta|\mathcal{D}) = \frac{1}{R}\sum_{r=1}^{R} p(\Delta|\boldsymbol{\beta}^r, \boldsymbol{\gamma}^r) \tag{4.15}$$

其中（$\boldsymbol{\beta}^r$，$\boldsymbol{\gamma}^r$）$\sim q_{\widehat{\boldsymbol{\eta}}}$（$\boldsymbol{\beta}$，$\boldsymbol{\gamma}$）。该过程在正式的贝叶斯框架下，同时考虑了模型结构 $\boldsymbol{\gamma}$和参数 $\boldsymbol{\beta}$的不确定性。这种方法的一个瓶颈在于，我们必须从参数和模型的庞大近似后验分布中进行采样，并且在预测阶段保留 $\widehat{\boldsymbol{\eta}}$ 的所有组成部分，这可能导致计算和内存效率低下。

基于后验均值的模型（Wasserman，2000年）在这种情况下，我们设 $\beta_{kj}^{(l)} = \hat{}E\{\beta_{kj}^{(l)}|\mathcal{D}\}$ 其中

$$E\{\beta_{kj}^{(l)}|\mathcal{D}\} = p(\gamma_{kj}^{(l)} = 1|\mathcal{D})E\{\beta_{kj}^{(l)}|\gamma_{kj}^{(l)} = 1, \mathcal{D}\} \approx \hat{\alpha}_{kj}^{(l)}\hat{\kappa}_{kj}^{(l)}.$$

此处，$\hat{}\alpha_{kj}^{(l)}$ 要么是通过变分推断程序获得的 $\alpha_{kj}^{(l)}$ 估计值，要么是在依赖结构 $\boldsymbol{\alpha}$ $E\{\alpha_{kj}^{(l)}|\mathcal{D}\}$的情况下得到的估计值。后一种情况下，理论上应当对 $\boldsymbol{\alpha}$ 积分消去，但这在实践中并不可行，因此避免了额外采样。该方法指定一个稠密模型$\hat{}\boldsymbol{\gamma}$，不进行稀疏化。同时，也无需采样。

中位概率模型（Barbieri 等人，2004）该方法基于中位概率模型这一概念；在简单线性模型的背景下，研究表明该模型在预测方面具有最优性。此处，我们设定$\gamma_{kj}^{(l)} = \mathrm{I}(\hat{}\alpha_{kj}^{(l)} > 0.5)$而 $\beta_{kj}^{(l)} \sim \gamma_{kj}^{(l)} N(\hat{\kappa}_{kj}^{(l)}, \hat{}\tau^2{}_{(l)kj})$。随后采用与（4.15）类似的模型平均方法。通过该方法，我们大幅稀疏化网络，并且只从边际纳入概率高于 0.5 的权重分布中进行采样。

结合参数后验均值的基于中位概率模型的推断 这里同样设定 $\gamma_{kj}^{(l)} = \mathrm{I}(\hat{}\alpha_{kj}^{(l)} > 0.5)$，但现在我们使用 $\beta_{kj}^{(l)} = \gamma_{kj}^{(l)}\hat{\kappa}_{kj}^{(l)}$。与基于后验均值的模型类似，这里无需进行采样。此外，我们只需存储 $\hat{\boldsymbol{\eta}}$ 中对应于边际纳入概率大于 0.5 的变分参数。因此，我们能够显著稀疏化所关注的贝叶斯神经网络（BNN），并大幅降低预测的计算成本。

后训练 一旦决定基于选定的模型进行推断，就可以在固定架构相关参数的情况下，针对模型的参数额外执行若干轮训练算法。这可能会进一步提高推断质量，同时也会使训练步骤变得更加容易，因为参数的数量大幅减少。这是因为不再需要估计边际纳入概率 $\boldsymbol{\alpha}$。此外，由于采用选定的中位概率模型所产生的稀疏性，与 $\gamma_{jk}^{(l)} = 1$ 相关的权重 $\beta_{jk}^{(l)}$ 的数量通常也会显著减少。还可以保持 $\alpha_{jk}^{(l)}$ 不变，同时仍允许 $\gamma_{jk}^{(l)}$ 保持随机。

其他模型选择准则和替代阈值化方法 中位概率模型（至少在理论上）并不总是可行的，因为需要在所有层之间至少存在一条连通路径，并且连接神经元的所有权重的边际纳入概率都高于 0.5。解决这一问题的一种方法是使用最可能模型（具有最大边际后验概率的模型）来替代中位概率模型。然后，在给定其配置的条件下，可以从参数分布中进行抽样，选择参数的均值（众数），或对参数的分布进行后训练。包括偏差信息准则和广泛适用信息准则在内的其他模型选择准则，也可以像最可能模型一样使用。处理这一问题的另一种启发式方法，是将以 $\mathrm{I}(\gamma_{kj}^{(l)} > 0.5)$ 为条件替换为以 $\mathrm{I}(\gamma_{kj}^{(l)} > \lambda)$ 为条件，其中 λ 是一个调节参数。在模型配置使用了过于保守的先验分布时，后一种方法也可能改善预测性能。同时，我们在实验中并未处理本段所述的方法，而是将其留待进一步研究。

5 应用

关于所建议的变分近似在线性回归模型背景下的深入研究，已在早期研究中开展，包括多个合成数据和真实数据示例，旨在恢复有意义的关系并进行预测（Carbonetto 等人，2012年；Hernández-Lobato 等人，2015年）。这些研究的结果表明，基于所建议的变分族分布的近似具有相当高的精确度，并且确实具备可扩展性，但可能存在偏差。本文将不讨论玩具示例和基于仿真的示例，而是请有兴趣的读者参阅上述参考文献中非常详尽且全面的研究；在本文中，我们将讨论一些更复杂的示例。具体而言，我们将讨论 MNIST（LeCun 等人，1998）和 Fashion-MNIST（FMNIST，Xiao 等人，2017年）图像的分类，以及 PHONEME 数据（Hastie 等人，1995年）。MNIST 和 FMNIST 数据集均包含 70000 张灰度图像（大小为 28x28），分为 10 个类别（分别为 0 到 9 的手写数字，以及 Zalando 的时尚商品“上衣”“裤子”“套头衫”“连衣裙”“外套”“凉鞋”“衬衫”“运动鞋”“包”和“踝靴”），每个类别包含 7 000 张图像。训练集包含 60000 张图像，测试集包含 10000 张图像。对于 PHONEME 数据集，响应中有 256 个协变量和 5 个类别。在该数据集中，训练集中有 3500 个观测值，测试集中有 1000 个观测值。PHONEME 数据取自 TIMIT 数据库（TIMIT 声学语音连续语料库，NTIS，美国商务部），该数据库是语音识别研究中广泛使用的资源。该数据集通过从该数据库的数字化语音中选取五个音素进行分类而构建。这些音素的转写如下：“sh”如同“she”中的发音，“dcl”如同“dark”中的发音，“iy”如同“she”中的元音，“aa”如同“dark”中的元音，“ao”如同“water”中的第一个元音。

实验设计 对于所有数据集，我们均采用使用 ReLU 激活函数的稠密神经网络，观测值服从多项分布。前两个示例包含 10 个类别和 784 个输入解释变量（像素），第三个示例则包含 256 个输入变量和 5 个类别。在这三种情况下，网络均有 2 个隐藏层，分别包含 400 和 600 个神经元。参数和模型指示变量的先验分布根据 (3.4) 选取，先验分布中的参数则通过经验贝叶斯方法指定。我们采用所提出的双重随机变分推断方法（算法 1）进行推断，训练 250 轮，批次大小为 100。M 设为 1，以降低计算成本；此外，选择该 M 也被认为足以与重参数化技巧结合使用（Gal，2016年）。最初的至多 20 个训练轮次用于对模型和参数进行预训练，并通过将先验分布的超参数（a_ψ、b_ψ、a_β、b_β）加入计算图，经验性地（亦称经验贝叶斯）估计这些超参数。之后，开始主要训练循环（先验分布的超参数保持固定）。我们使用 ADAM 随机梯度上升优化（Kingma 和 Ba，2014年），在预训练阶段和主要训练阶段采用算法 1 中的对角矩阵 $\boldsymbol{A}$，其对角元素见表 1 和表 S-1。通常，经验贝叶斯方法会最大化边际似然；但由于我们无法针对所研究的模型计算边际似然，因此在预训练阶段使用其下界（证据下界，ELBO）。经过 250 个训练轮次后，我们进行了后训练。对参数进行后训练时，无论采用固定的边际纳入概率，还是采用中位概率模型，我们都使用表 1 和表 S-1 底部各行中指定的 $\boldsymbol{A}$，额外运行 50 个轮次的优化程序。对于完全贝叶斯模型平均方法，我们同时使用了 $R = 1$ 和 $R = 10$。尽管 $R = 1$ 可能无法准确估计预测分布的蒙特卡罗估计，但由于其稀疏化程度较高，仍值得关注。实验中使用的所有 PyTorch 实现均可在我们的 GitHub 代码库中获取。

我们报告了将我们的模型 **LBBNN** 应用于 β's 的 slab 成分采用高斯先验（**GP**），并基于平均场（**MF**）、多元正态分布（**MVN**）和低秩多元正态分布（**LFMVN**；其预测结果见补充材料）变分推断潜在指标之间依赖结构的结果。我们分别使用组合名称**LBBNN-GP-MF**、**LBBNN-GP-MVN**和**LBBNN-GP-LFMVN** 表示模型、先验分布和变分分布的组合。论文补充材料第 4 节报告了结果

表 1： LBBNN-GP-MF 和 LBBNN-GP-MVN 的优化算法步长矩阵 $\boldsymbol{A}$ 的对角元素规格，缩写说明见表 2 。请注意， A_ω 仅用于 LBBNN-GP-MF，而 A_ξ 和 A_Σ 仅用于 LBBNN-GP-MVN。LBBNN-GP-LFMVN 的调优参数见论文补充材料中的表 S-1。

	A_β, A_ρ	A_ξ	A_ω	A_Σ	A_{a_ψ},A_{b_ψ}	A_{a_β},A_{b_β}
预训练	0.00010	0.10000	0.10000	0.10000	0.00100	0.00001
训练	0.00010	0.01000	0.00010	0.00010	0.00000	0.00000
后训练	0.00010	0.00000	0.00000	0.00000	0.00000	0.00000

我们早期预印本（Hubin 和 Storvik，2019）中关于 MNIST 和 FMNIST 数据集的结果；这些数据集的超参数固定（对应于 Bai 等人，2020）所考虑的设置）。

表 2： 缩写 实验结果报告中使用的设置和评估指标 实验结果。

模型	**含义**
BNN	贝叶斯神经网络
LBBNN	潜在二元贝叶斯神经网络
参数先验分布	**含义**
GP	权重的独立高斯先验
MGP	权重的独立高斯混合先验
HP	权重的独立马蹄先验
推断	**含义**
MF	平均场变分推断
MVN	纳入概率的多元高斯结构
LFMVN	纳入概率多元正态分布结构的协方差低秩因子
γ	**含义**
SIM	权重的纳入情况根据纳入指示变量的后验分布抽取
ALL	使用全部权重
MED	使用与中位概率模型对应的权重
PRN	未使用基于阈值的规则进行剪枝，使用全部权重
β	**含义**
SIM	所包含的权重取自其后验分布
MEA	使用权重的后验均值
R	**含义**
10	抽取 10 个样本
1	抽取 1 个样本或使用后验均值
评估指标	**含义**
所有类别准确率	针对测试集中的所有样本计算的准确率
0.95阈值准确率	针对测试集中最大值超过阈值的样本计算的准确率
0.95 阈值	（跨类别）模型平均预测后验概率超过 0.95 测试集中最大值
类别数	（跨类别）模型平均预测后验概率超过 0.95
密度级别	用于进行预测的权重占比
每轮时间	每个训练轮次的平均耗时

此外，我们还采用了若干相关的 基线方法。具体而言，我们研究了一种采用高斯先验和平均场变分推断的标准稠密贝叶斯神经网络（Graves，2011年），记为 **BNN-GP-MF**，它可视为原始模型的一种特例，其中所有 $\gamma_{kj}^{(l)}$ 均固定为 1，且权重的方差分量没有先验分布。该模型对于衡量引入稀疏性后预测能力的变化十分重要。此外，我们还报告了采用混合先验的稠密贝叶斯神经网络（**BNN-MGP-MF**）的结果，该模型使用两个高斯分量构成混合分布（Blundell 等人，2015年），各分量的概率均为 0.5，方差分别为 1 和 e^{-6} 。此外，我们还研究了两种常用的稀疏化方法，具体而言，分别是采用 Concrete dropout 的稠密网络（**BNN-GP-CMF**）（Gal 等人，2017年）和采用马蹄先验的稠密网络（**BNN-HP-MF**）（Louizos 等人，2017年）。最后，我们采用频率学派全连接神经网络（**前馈神经网络**）（经事后权重剪枝）作为更基础的基线方法。为使实验设计更简洁，我们仅在补充材料中报告前馈神经网络的结果。所有基线方法（包括前馈神经网络）也都具有 2 个隐藏层，分别包含 400 和 600 个神经元。它们采用 Adam 优化器训练了 250 个训练轮次（所有相关参数的学习率为 $a = 0.0001$ ），批次大小为 100。对于采用马蹄先验的贝叶斯神经网络，我们分别报告权重临时剪枝（PRN）前后的统计数据。后训练（如有需要）额外进行了 50 个训练轮次。对于前馈神经网络，在三个实验中，我们都进行了权重和神经元剪枝（Blalock 等人，2020），使其稀疏度水平与贝叶斯方法获得的结果相同，以便直接比较。前馈神经网络的剪枝基于移除相应比例中幅值（绝对值）最小的权重/神经元。前馈神经网络未考虑不确定性，也未考虑结构学习。

关于预测，第4.3节介绍了几种方法。这些方法本质上都归结为如何处理模型参数 $\boldsymbol{\gamma}$ 和权重 $\boldsymbol{\beta}$的选择。对于 $\boldsymbol{\gamma}$，我们可以从（近似）后验分布中进行模拟（**SIM**），也可以使用中位概率模型（**MED**）。此外，针对**BNN-HP-MF** ，还可以采用**PRN**剪枝方法，该方法由Louizos 等人(2017年)提出。我们还考虑了对一些基线方法采用包含全部权重的方案（**全部**）。对于$\boldsymbol{\beta}$，我们考虑从（近似）后验分布中采样（**SIM**），或使用后验均值（**MEA**）。按照这一记法，第4.3 节中的完全贝叶斯模型平均记为**SIM SIM**，基于后验均值的模型记为**ALL MEA**，中位概率模型记为**MED SIM**，而中位概率模型与参数后验均值结合的方案记为**MED MEA**。

随后，我们评估了准确率（Acc，即正确分类图像所占的比例）。我们还计算了基于中位概率模型（通过 $R = 1$ 或 $R = 10$）以及后验均值模型的准确率。最后，我们评估了在边际纳入概率固定的情况下对参数进行后训练，以及对中位概率模型进行后训练后得到的准确率。对于涉及模型平均的情况（$R = 10$），我们还报告了仅当模型平均后的最大类别概率超过 95% 时才进行分类所得到的准确率，这一做法由 Posch等（2019）提出。否则，则作出存疑判定（Ripley，2007，第 2.1 节）。在这种情况下，我们既报告已分类图像中的准确率，也报告已分类图像的数量。最后，我们报告不同方法下的总体密度水平（至少在一次仿真中 $\gamma_{kj}^{(l)}$ 等于 1 的比例）。为确保结果可复现，我们对所述实验 $s \in \{1, ..., 10\}$ 的 10 次独立运行所得的所有这些统计量计算了汇总值（中位数、最小值和最大值）。此外，我们还根据所建议的变分近似，计算了所有权重的边际纳入概率估计值 $\hat{}p(\gamma_{kj}^{(l)} = 1|\mathcal{D})$。为精简结果呈现，我们仅列出各层 l 的平均边际纳入概率，记为 $\rho(\gamma^{(l)}|\mathcal{D}) := \frac{1}{p^{(l+1)}p^{(l)}} \sum_{kj} \hat{}p(\gamma_{kj}^{(l)} = 1|\mathcal{D})$，汇总于表6中。为便于理解结果中使用的缩写，我们在表 2 中列出了相应的简要说明。

MNIST 表 3 和 6 中报告的结果（补充材料的表 S-2 和 S-8 中还报告了 LBBNN-GP-LFMVN 和后训练的一些额外结果）表明，在我们的 LBBNN 方法中：a) 模型平均

不同贝叶斯神经网络之间的$R = 10$）准确率显著高于从模型空间中随机抽取的单个贝叶斯神经网络的准确率（$R = 1$）；b) 中位概率模型和基于后验均值的模型的表现也显著优于随机抽取的模型。中位概率模型和基于后验均值的模型的表现实际上与完整模型平均相当；c) 根据表 6 和图 1，对于平均场变分分布，各模型中第1层和第2层权重的大多数边际纳入概率都很低，而第3层有更多权重的边际纳入概率较高（尽管该层的权重也显著减少）。这与卷积神经网络（CNN）的结构相似：通常先是若干稀疏卷积层，随后是几个全连接层。与 CNN 不同的是，我们的方法能够自动学习稀疏化结构；d) 对于采用全秩结构的多元正态分布变分近似，输入层最为稠密，其次是第二层的极度稀疏化，以及第3层的中度稀疏化；e) 采用低阶协方差矩阵因子参数化的多元正态分布方法（结果见补充材料）只能实现非常有限的稀疏化，权重参数的稀疏化比例不超过 50%；f) 各项性能指标在不同仿真中的变化都很小，表明重复实验的结果稳定；g) 采用弃权选项进行推断时，准确率几乎完美，但代价是拒绝对部分样本进行分类。

对于其他方法，结论也相同：h) 与参数空间中的单个样本相比，使用基于后验均值的模型和使用样本平均都能提高准确率；i) 对于采用高斯先验/高斯混合先验的稠密贝叶斯神经网络，以及采用马蹄先验的贝叶斯神经网络，目标参数估计值的变异性较低，而 Concrete dropout 方法的变异性相当高。在将我们的方法与基线方法进行比较时，我们发现：j) 总体而言，稠密方法的准确率优于稀疏方法；k) Concrete dropout 的中位准确率略高于其他方法，但其方差较大，而我们的完整贝叶斯神经网络和采用马蹄先验的压缩贝叶斯神经网络在各项实验中表现稳定；l) 无论是我们的方法还是基线方法，在预测硬分类准确率方面都未能达到最先进水平（Palvanov 和 Im Cho，2018）；m) 将 95%阈值用于分类，会使采用马蹄先验时被分类的案例数极少（其置信度严重不足）；采用弃权选项进行推断时，Concrete dropout 方法似乎过度自信（导致准确率较低，但决策数量较多）；完整贝叶斯神经网络，以及采用高斯先验和高斯混合先验的贝叶斯神经网络，分类案例数少于 Concrete dropout 方法，但准确率显著更高；n) 这可能意味着需要针对特定方法校准阈值；o) 使用中位概率模型时，我们的方法在平均场变分近似和全秩多元正态结构的变分近似下，权重稀疏性最高。此外，q) 后训练（结果见补充材料）似乎并未显著提升模型的预测质量或不确定性处理能力；o) 在给定网络深度和宽度配置下，对于所有考虑的稀疏度水平，贝叶斯神经网络在泛化误差方面均显著优于具有相应相同稀疏度水平的频率学派对应方法。最后，就计算时间而言，正如预期，前馈神经网络的每轮训练用时最短；对于贝叶斯方法，参数数量与计算时间呈显著正相关，其中 BNN-GP-CMF 是最快的方法，LBBNN-GP-MVN 则最慢。所有时间数据均是在 GeForce RTX 2080 Ti GPU 上训练模型时获得的。尽管如此，值得注意的是，最快与最慢的贝叶斯方法之间的速度差异不到 3 倍；考虑到时间还会受到不同方法的实现方式以及实验运行时服务器负载可能不同的影响，这种差异在实践中完全可以接受。

FMNIST 此应用采用了与 MNIST 示例相同的一组方法、模型规格和算法调优参数。根据补充材料中的表4、6以及表 S-3 和 S-9，以及图1 ，FMNIST 数据的结果 a)- r) 与 MNIST 实验的结果完全一致；不过，所有方法在 FMNIST 上的预测性能都较差。此外，尽管完整贝叶斯神经网络和采用马蹄先验的贝叶斯神经网络在 FMNIST 上的稀疏度水平低于在 MNIST 上的水平，Concrete dropout 在此处有所改善

表 3：比较方法在 MNIST 数据上的性能指标。所有结果均为 10 次重复实验的中位数（括号内包含最小值和最大值）。未使用后训练。详情请参见表 2。

预测 γ	β	模型先验-方法	R	所有类别 Acc	0.95 阈值 Acc	0.95 阈值 类别数	密度级别	Epo. time
SIM	SIM	LBBNN-GP-MF	1	0.968（0.966，0.970）	-	-	0.090	8.363
SIM	SIM	LBBNN-GP-MF	10	0.981 (0.979,0.982)	0.999	8322	1.000	8.363
ALL	MEA	LBBNN-GP-MF	1	0.981 (0.980,0.983)	-	-	1.000	8.363
MED	SIM	LBBNN-GP-MF	1	0.969（0.968，0.974）	-	-	0.079	8.363
MED	SIM	LBBNN-GP-MF	10	0.980 (0.979,0.982)	0.999	8444	0.079	8.363
MED	MEA	LBBNN-GP-MF	1	0.981 (0.980,0.983)	-	-	0.079	8.363
SIM	SIM	LBBNN-GP-MVN	1	0.965 (0.964,0.966)	-	-	0.180	9.651
SIM	SIM	LBBNN-GP-MVN	10	0.978（0.976,0.979）	1.000	7818	1.000	9.651
ALL	MEA	LBBNN-GP-MVN	1	0.978（0.976，0.980）	-	-	1.000	9.651
MED	SIM	LBBNN-GP-MVN	1	0.968（0.966，0.969）	-	-	0.163	9.651
MED	SIM	LBBNN-GP-MVN	10	0.977（0.975,0.979）	1.000	7928	0.163	9.651
MED	MEA	LBBNN-GP-MVN	1	0.974（0.972，0.976）	-	-	0.163	9.651
ALL	SIM	BNN-GP-MF	1	0.965 (0.965,0.966)	-	-	1.000	5.094
ALL	SIM	BNN-GP-MF	10	0.984 (0.982,0.985)	0.999	8477	1.000	5.094
ALL	MEA	BNN-GP-MF	1	0.984（0.982,0.985）	-	-	1.000	5.094
ALL	SIM	BNN-MGP-MF	1	0.965 (0.964,0.967)	-	-	1.000	5.422
ALL	SIM	BNN-MGP-MF	10	0.982 (0.981,0.983)	0.999	8329	1.000	5.422
ALL	MEA	BNN-MGP-MF	1	0.983（0.981，0.984）	-	-	1.000	5.422
SIM	SIM	BNN-GP-CMF	1	0.982（0.894，0.984）	-	-	0.226	3.477
SIM	SIM	BNN-GP-CMF	10	0.984（0.896，0.986）	0.995	9581	1.000	3.477
ALL	MEA	BNN-GP-CMF	1	0.984（0.893，0.986）	-	-	1.000	3.477
SIM	SIM	BNN-HP-MF	1	0.964 (0.962,0.967)	-	-	1.000	4.254
SIM	SIM	BNN-HP-MF	10	0.982 (0.981,0.983)	1.000	0003	1.000	4.254
ALL	MEA	BNN-HP-MF	1	0.966（0.963，0.968）	-	-	1.000	4.254
PRN	SIM	BNN-HP-MF	1	0.965 (0.962,0.969)	-	-	0.194	4.254
PRN	SIM	BNN-HP-MF	10	0.982 (0.981,0.983)	1.000	0002	0.194	4.254
PRN	MEA	BNN-HP-MF	1	0.965（0.963，0.968）	-	-	0.194	4.254

与前一个示例相比，这方面有所不同。对于前馈神经网络，针对 MNIST 数据集得出的相同结论仍然成立。

表 4： 本文提出的 BNN 贝叶斯方法在 FMNIST 数据上的性能指标。更多详情请参见表 2 以及表3的标题。

预测 γ	β	模型先验-方法	R	所有类别 Acc	0.95 阈值 Acc	0.95 阈值 类别数	密度级别	Epo. time
SIM	SIM	LBBNN-GP-MF	1	0.864（0.861，0.866）	-	-	0.120	7.969
SIM	SIM	LBBNN-GP-MF	10	0.883（0.881，0.886）	0.995	4946	1.000	7.969
ALL	MEA	LBBNN-GP-MF	1	0.882（0.879，0.887）	-	-	1.000	7.969
MED	SIM	LBBNN-GP-MF	1	0.867 (0.864,0.871)	-	-	0.108	7.969
MED	SIM	LBBNN-GP-MF	10	0.883（0.880，0.886）	0.995	5025	0.108	7.969
MED	MEA	LBBNN-GP-MF	1	0.880（0.877，0.886）	-	-	0.108	7.969
SIM	SIM	LBBNN-GP-MVN	1	0.858（0.854，0.859）	-	-	0.156	9.504
SIM	SIM	LBBNN-GP-MVN	10	0.879（0.874，0.880）	0.995	4503	1.000	9.504
ALL	MEA	LBBNN-GP-MVN	1	0.875（0.873，0.876）	-	-	1.000	9.504
MED	SIM	LBBNN-GP-MVN	1	0.865 (0.860,0.866)	-	-	0.129	9.504
MED	SIM	LBBNN-GP-MVN	10	0.877 (0.875,0.879)	0.995	4694	0.129	9.504
MED	MEA	LBBNN-GP-MVN	1	0.871（0.868，0.875）	-	-	0.129	9.504
ALL	SIM	贝叶斯神经网络-GP-平均场	1	0.864 (0.863,0.866)	-	-	1.000	5.368
ALL	SIM	BNN-GP-MF	10	0.893（0.890，0.894）	0.997	5089	1.000	5.368
ALL	MEA	BNN-GP-MF	1	0.886（0.882，0.888）	-	-	1.000	5.368
ALL	SIM	BNN-MGP-MF	1	0.867（0.866,0.868）	-	-	1.000	4.803
ALL	SIM	BNN-MGP-MF	10	0.893 (0.892,0.897)	0.996	5151	1.000	4.803
ALL	MEA	BNN-MGP-MF	1	0.888 (0.885,0.890)	-	-	1.000	4.803
SIM	SIM	BNN-GP-CMF	1	0.896 (0.820, 0.902)	-	-	0.094	3.369
SIM	SIM	BNN-GP-CMF	10	0.897（0.823，0.901）	0.942	8825	1.000	3.369
ALL	MEA	BNN-GP-CMF	1	0.896 (0.821,0.901)	-	-	1.000	3.369
SIM	SIM	BNN-HP-MF	1	0.864（0.863，0.869）	-	-	1.000	4.613
SIM	SIM	BNN-HP-MF	10	0.887 (0.886,0.889)	1.000	0181	1.000	4.613
ALL	MEA	BNN-HP-MF	1	0.867 (0.861, 0.868)	-	-	1.000	4.613
PRN	SIM	BNN-HP-MF	1	0.865（0.860，0.868）	-	-	0.302	4.613
PRN	SIM	BNN-HP-MF	10	0.887（0.884，0.888）	1.000	0179	0.302	4.613
PRN	MEA	BNN-HP-MF	1	0.865（0.862，0.869）	-	-	0.302	4.613

PHONEME 最后，PHONEME 数据的分类也采用了与 MNIST 和 FMNIST 示例相同的一组方法、模型规格（输入协变量为 256 个、响应类别为 5 类除外）以及算法调优参数。根据补充材料中的表 5、6 以及表 S-4 和 S-10，PHONEME 数据的结果 a)–r) 总体上也与 MNIST 和 FMNIST 实验的结果一致。不过，所有方法的预测性能都优于 FMNIST，但不及 MNIST。对于所有可以进行稀疏化的方法，本例的稀疏性级别都较低。不过，结果仍表明实现相当程度的稀疏化是可行的。对于前馈神经网络（FNN），MNIST 和 FMNIST 数据集上的结论同样适用，不过这里 FNN 性能的下降没有那么明显。此外，正如补充材料中的图 S-5–S-8 所示，这些结论在贝叶斯神经网络的各种宽度配置下均保持一致

表5： 本文提出的 BNN 贝叶斯方法在 PHONEME 数据上的性能指标。更多详情请参见表 2 及表 3的标题。

预测 γ	β	模型先验-方法	R	所有类别 Acc	0.95 阈值 Acc	0.95 阈值 类别数	密度级别	Epo. time
SIM	SIM	LBBNN-GP-MF	1	0.913（0.898，0.929）	-	-	0.371	0.433
SIM	SIM	LBBNN-GP-MF	10	0.927（0.923，0.933）	0.992	690	1.000	0.433
ALL	MEA	LBBNN-GP-MF	1	0.925 (0.921,0.933)	-	-	1.000	0.433
MED	SIM	LBBNN-GP-MF	1	0.923 (0.910,0.928)	-	-	0.307	0.433
MED	SIM	LBBNN-GP-MF	10	0.925（0.912,0.934）	0.984	757	0.307	0.433
MED	MEA	LBBNN-GP-MF	1	0.925 (0.913,0.932)	-	-	0.307	0.433
SIM	SIM	LBBNN-GP-MVN	1	0.919 (0.911,0.927)	-	-	0.255	0.505
SIM	SIM	LBBNN-GP-MVN	10	0.929 (0.927,0.935)	0.995	649	1.000	0.505
ALL	MEA	LBBNN-GP-MVN	1	0.926（0.918，0.931）	-	-	1.000	0.505
MED	SIM	LBBNN-GP-MVN	1	0.925 (0.916,0.929)	-	-	0.225	0.505
MED	SIM	LBBNN-GP-MVN	10	0.929 (0.925,0.933)	0.995	668	0.225	0.505
MED	MEA	LBBNN-GP-MVN	1	0.924 (0.921,0.928)	-	-	0.225	0.505
ALL	SIM	BNN-GP-MF	1	0.915（0.907，0.919）	-	-	1.000	0.203
ALL	SIM	BNN-GP-MF	10	0.919 (0.900,0.929)	0.966	834	1.000	0.203
ALL	MEA	BNN-GP-MF	1	0.917（0.901，0.922）	-	-	1.000	0.203
ALL	SIM	BNN-MGP-MF	1	0.913（0.910，0.925）	-	-	1.000	0.208
ALL	SIM	BNN-MGP-MF	10	0.916 (0.912,0.926)	0.969	833	1.000	0.208
ALL	MEA	BNN-MGP-MF	1	0.921 (0.914,0.926)	-	-	1.000	0.208
SIM	SIM	BNN-GP-CMF	1	0.879（0.706，0.906）	-	-	0.509	0.103
SIM	SIM	BNN-GP-CMF	10	0.922 (0.918,0.930)	0.965	187	1.000	0.103
ALL	MEA	BNN-GP-CMF	1	0.873（0.712，0.904）	-	-	1.000	0.103
SIM	SIM	BNN-HP-MF	1	0.921 (0.915,0.929)	-	-	1.000	0.136
SIM	SIM	BNN-HP-MF	10	0.921 (0.915,0.926)	0.895	019	1.000	0.136
ALL	MEA	BNN-HP-MF	1	0.921（0.916，0.926）	-	-	1.000	0.136
PRN	SIM	BNN-HP-MF	1	0.919 (0.909,0.926)	-	-	0.457	0.136
PRN	SIM	BNN-HP-MF	10	0.919（0.916，0.927）	0.926	028	0.457	0.136
PRN	MEA	BNN-HP-MF	1	0.920（0.914，0.926）	-	-	0.457	0.136

。此外，对于所有方法，随着宽度增加，稀疏化程度也会提高，但稀疏性的增长与宽度的增长并不成正比。

域外实验 遵循Nitarshan (2018）提出的测量域内和域外不确定性的方法示例，我们首先考察 LBBNN-GP-MF 方法通过使用基于在 MNIST 数据集上训练的模型和参数的联合后验分布所得到的后验预测分布样本，尝试对来自 FMNIST 图像的样本进行分类，从而判断其预测置信度的能力，并将其与 MNIST 数据测试集中的图像样本的结果进行比较。这里报告的是实验运行 $s = 10$ 所获得的联合后验分布（模型和参数）的结果。如图 2 所示，对于 MNIST 数据集，LBBNN-GP-MF 给出的预测具有很高的置信度，后验预测分布中的样本几乎没有方差。与此同时，与基于 FMNIST 数据的后验预测分布样本相关的域外不确定性通常较高（存在一些例外），这表明来自

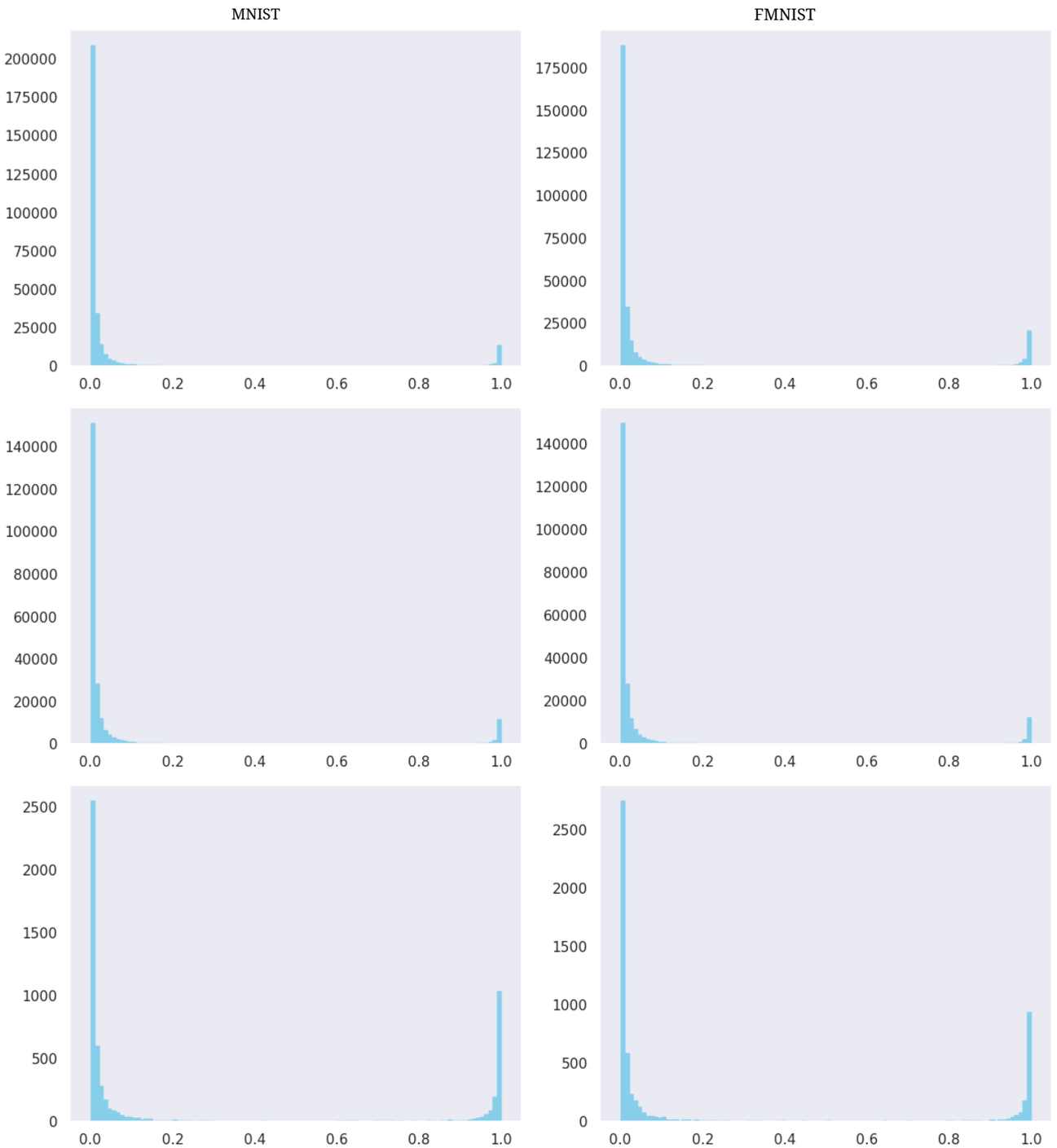

图1：LBBNN-GP-MF 的三个层（从上到下）权重边际纳入概率直方图示意图，来自 MNIST（左）和 FMNIST（右）的仿真 $s = 10$ 。

表6：在 MNIST 和 FMNIST 数据上进行的 10 次重复实验中，我们模型的边际纳入概率（每层取平均；定义见正文）的中位数和标准差。

	MNIST 数据	FMNIST 数据	PHONEME 数据
LBBNN-GP-MF			
$\rho(\gamma^{(1)}=1\|\mathcal{D})$	0.0844 (0.0835,0.0853)	0.1323 (0.1291,0.1349)	0.3806 (0.3764,0.3838)
$\rho(\gamma^{(2)}=1\|\mathcal{D})$	0.0959 (0.0942, 0.0967)	0.1005 (0.0981,0.1020)	0.3670（0.3641，0.3699）
$\rho(\gamma^{(3)}=1\|\mathcal{D})$	0.2945 (0.2808,0.3056)	0.2790 (0.2709,0.2921)	0.4236（0.4053，0.4367）
LBBNN-GP-MVN			
$\rho(\gamma^{(1)}=1\|\mathcal{D})$	0.2975 (0.2928,0.2993)	0.2461 (0.2410,0.2515)	0.3201（0.3142,0.3273）
$\rho(\gamma^{(2)}=1\|\mathcal{D})$	0.0368 (0.0363,0.0377)	0.0392 (0.0383,0.0398)	0.2287 (0.2235,0.2355)
$\rho(\gamma^{(3)}=1\|\mathcal{D})$	0.1394 (0.1311,0.1475)	0.1462 (0.1368,0.1521)	0.2763（0.2632，0.2953）
LBBNN-GP-LFMVN			
$\rho\ (\gamma^{(1)}=1\|\mathcal{D})$	0.4474（0.4448，0.4498）	0.4589（0.4565，0.4603）	0.4973（0.4965，0.4987）
$\rho(\gamma^{(2)}=1\|\mathcal{D})$	0.4525 (0.4501,0.4537)	0.4516 (0.4493,0.4528)	0.4972 (0.4952,0.4990)
$\rho(\gamma^{(3)}=1\|\mathcal{D})$	0.4815 (0.4685,0.4871)	0.4805 (0.4654,0.4868)	0.4979 (0.4925,0.5048)

此处的后验预测分布。关于 FMNIST 推断和 MNIST 数据相关不确定性的逆向示例（如图3所示）得出了相同的结论。LBBNN-GP-MVN 和 LBBNN-GP-LFMVN 方法得到的结果也大致相同，但由于篇幅限制，论文中未予报告。

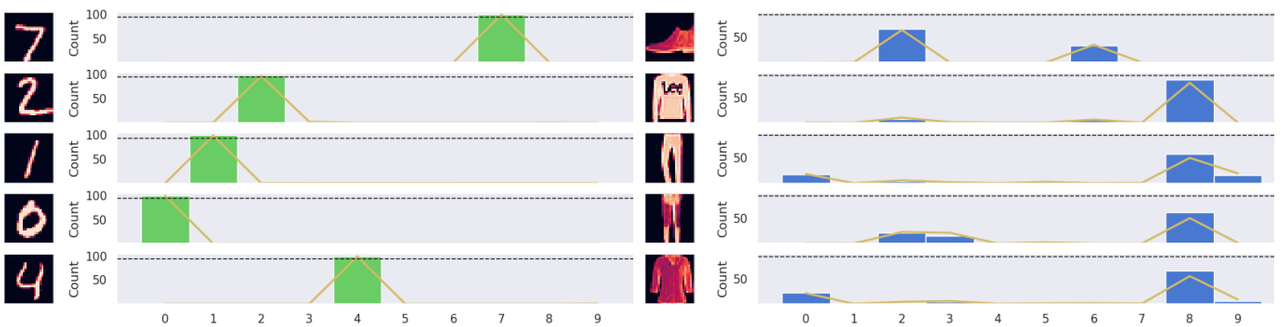

图2：基于后验预测分布中的 100 个样本，展示域内测试数据（MNIST，左）和域外测试数据（FMNIST，右）相关的不确定性。黄色线表示模型平均后验类别概率（百分比）。绿色条标示正确类别，蓝色条表示其他样本（条形高度对应于预测过程各次重复中采用硬分类得到的类别概率的另一种估计值）。黑色虚线表示做出决策时的 95%阈值，可据此判断是否存在疑虑）。原始图像显示在左侧。

图 4展示了更详细的域外实验结果：使用 FMNIST 数据测试在 MNIST 数据上训练的模型，反之亦然。按照 Louizos 和 Welling (2017年)，现在的目标是尽可能得到不确定的结果（理想情况下，各类别的概率均匀分布），这对应较大的熵。图中展示了已分类样本的熵的经验累积分布函数（CDF）；理想情况下，CDF 应接近右下角。Concrete dropout 的预测过度自信，其测试类别分布远非均匀；基于马蹄先验的方法（剪枝前后）最接近均匀分布（但它在域内预测中也更接近均匀分布），而它

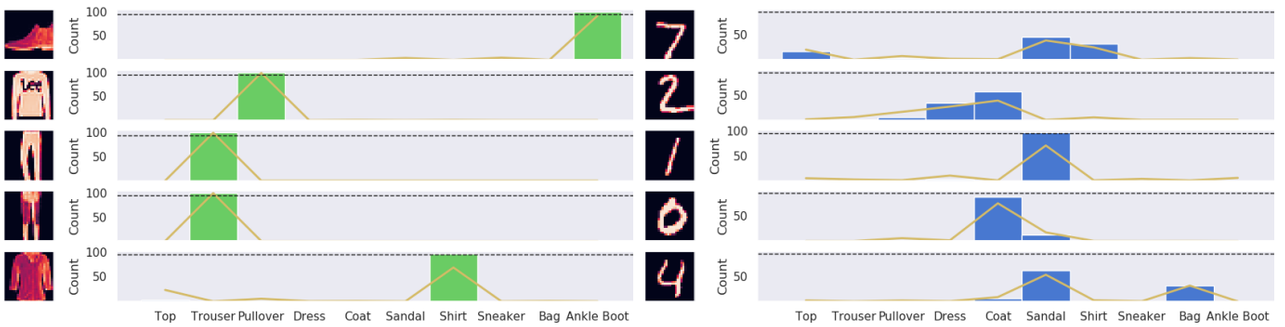

图3：基于后验预测分布中的100个样本，与域内测试数据（FMNIST，左）和域外测试数据（MNIST，右）相关的不确定性。更多细节请参见图2 。

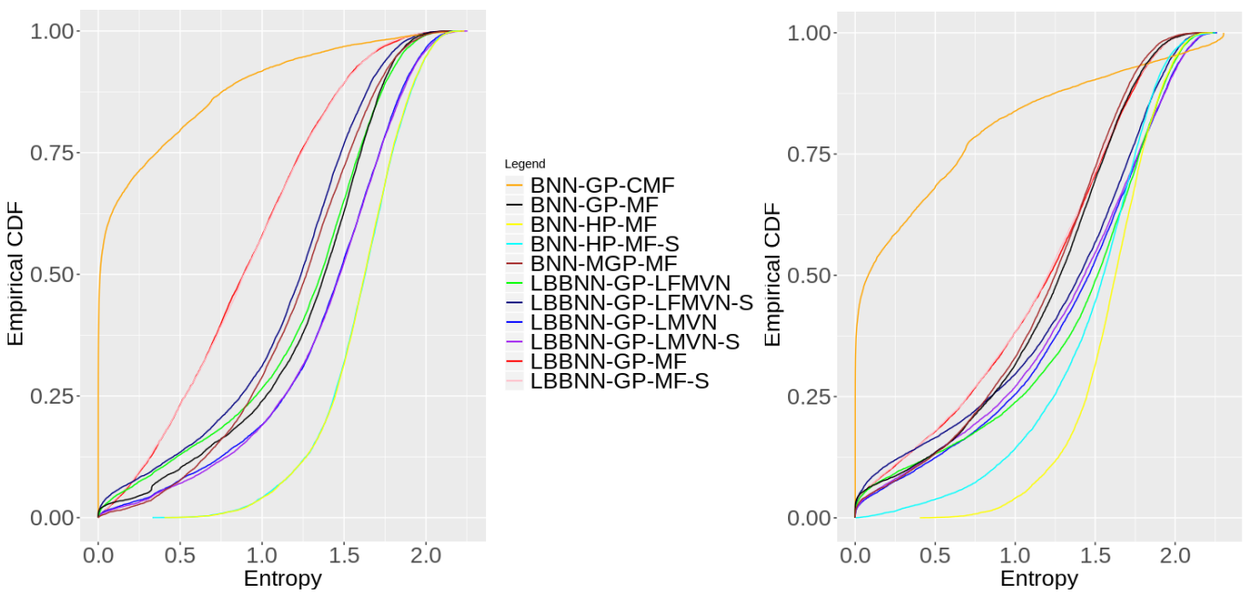

图4：在 MNIST 上训练并应用于 FMNIST（左），以及反之（右）的边缘后验预测分布熵的经验累积分布函数，对应仿真 $s = 10$。后缀 S 表示通过适当方法进行稀疏化的模型。

也更接近均匀分布；另外两种基线方法介于两者之间。我们的方法（剪枝前，以及采用中位概率模型剪枝后）与它们表现相当，说明我们能够较好地处理域外不确定性。

关于误分类不确定性的更多内容 图 5 展示了与后验预测抽样相关的误分类不确定性。可以看出，在 LBBNN-GP-MF 发生误分类的大多数情况下，预测的类别确定性相对较低，表明网络并不确定。此外，即使在这些情况下，真实类别通常也位于预测的 95%可信区间内。根据 Posch 等人(2019) 的说法，可以通过以下方式判断：100 个样本中属于错误类别的少于 95 个，且至少有 6 个样本属于正确类别。还要注意，在此处展示的许多误分类案例中，即使是人类也会对如何作出决定感到十分犹豫。这里，LBBNN-GP-MVN 和 LBBNN-GP-LFMVN 方法同样得到了非常相似的结果，但由于篇幅所限，论文中未予报告。

6 讨论

本文介绍了贝叶斯神经网络中的贝叶斯模型（或结构）不确定性概念，并提出了一种可扩展的变分推断技术，用于近似模型及其参数的联合后验分布。通过对模型和参数进行边缘化，可以轻松获得近似后验预测分布。此外，边际纳入概率为贝叶斯二元丢弃提供了恰当的概率解释，并可用于模型（或架构）选择。为此，每个纳入的权重只需增加一个参数。

我们通过图像和声音分类应用展示了所提技术的效果：它既能在不明显损失预测能力的情况下显著稀疏化神经网络，也能准确处理预测不确定性。关于优化的计算成本：对于平均场近似，我们为每个权重仅增加一个参数 α_{kj}^l 。在 $\boldsymbol{\alpha}^{(l)}$上采用底层高斯结构时，还会进一步引入协方差矩阵的附加参数。每个优化步骤的复杂度与待优化参数的数量成正比，因此，与全连接贝叶斯神经网络甚至前馈神经网络相比，计算时间的增加（如实验所示）并不显著。对于所获得的预测，不同方法的复杂度与参与预测的“活跃”参数数量成正比，因此通常有利于更稀疏的方法。

关于实践建议，根据我们的实证结果，如果希望在稀疏性、预测准确率、不确定性以及计算成本之间取得合理的权衡，我们建议使用 LBBNN-GP-MF。如果不需要稀疏性，标准的 BNN-GP-MF 和 BNN-MGP-MF 就足够了。

目前，对模型和参数均采用相当简单的先验分布。这些先验分布被假定在神经网络的各个参数之间相互独立，但这种假设并不总是合理的。另一种做法是，让参数先验和模型先验都包含联合依赖结构，从而进一步提升神经网络配置的稀疏化效果。对于具有局部结构以及变量（神经元）间依赖关系的模型先验，可以举出所谓的稀释先验（George 等人，2010）。这类先验通过降低含有高度相关变量的模型的概率，来处理模型间的相似性。此外，在更简单的模型中，也有许多方法可在不同设定下通过先验将模型参数之间的相互依赖关系纳入考量（Smith 和 LeSage，2004；Fahrmeir 和 Lang，2001；Dobra 等人，2004）。显然，在贝叶斯神经网络中对参数与模型进行联合推断的背景下，先验分布的选择仍需进一步研究。具体而言，对于图像分析，开发卷积诱导先验可能会很有意义；对于循环模型，则可以考虑采用指数衰减的参数先验来控制短期与长期记忆。

在本研究中，我们仅考虑贝叶斯神经网络的一个子类，该子类通过在给定架构中纳入或排除特定权重来定义。未来，将该方法扩展到选择

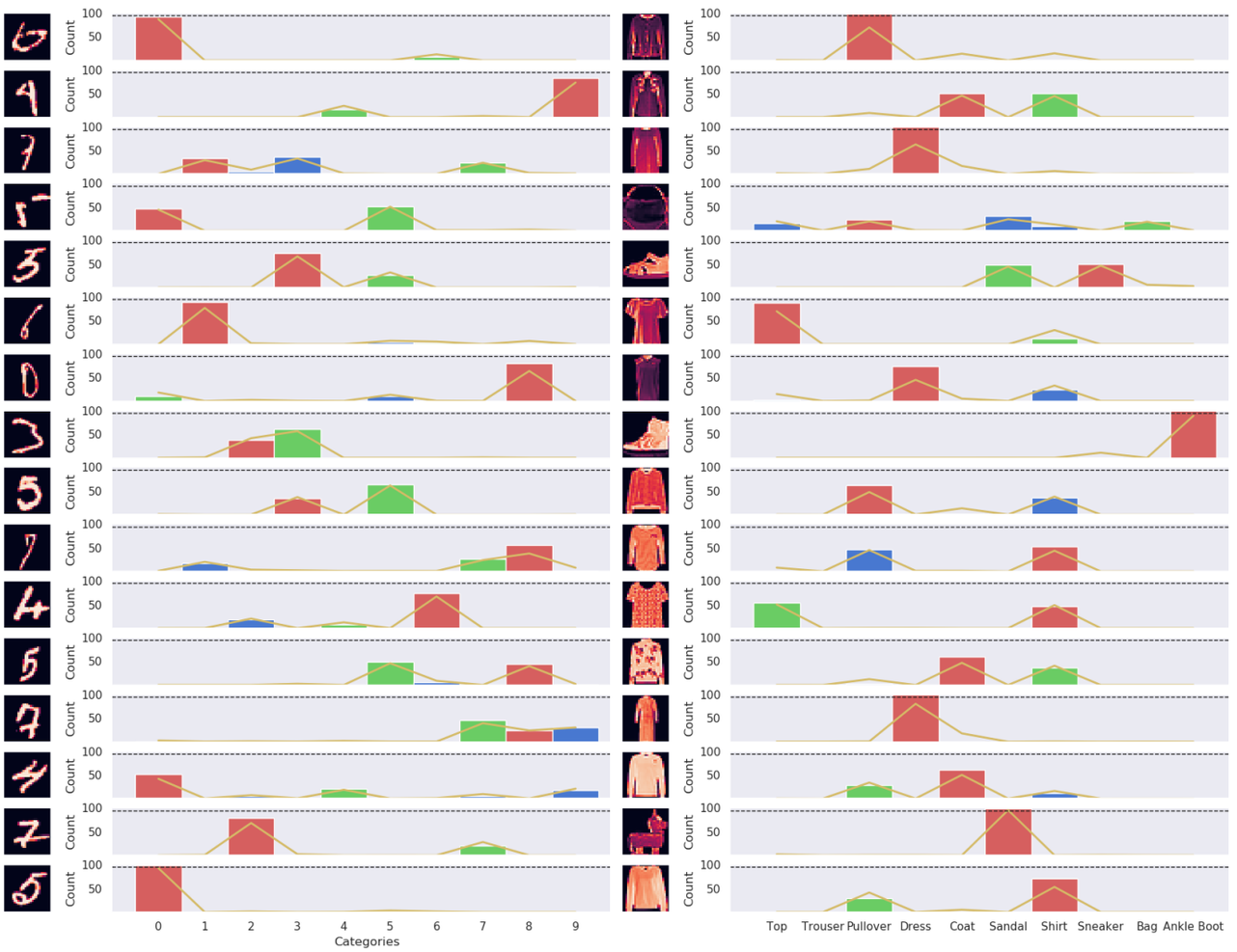

图5：基于 LBBNN-GP-MF 模型联合后验分布（来自仿真$s = 10$）样本的不确定性，针对 16 张可能被错误分类（在模型平均下）的 MNIST 数据（左）和 FMNIST 数据（右）图像。黄色线条表示基于完整贝叶斯神经网络方法的模型平均后验类别概率（百分比）。绿色条表示真实类别，红色条表示错误预测的类别，蓝色条表示其他样本；黑色虚线表示包含疑问选项时做出决策的 95%阈值。原始图像显示在左侧。

贝叶斯神经网络的激活函数，以及每一层的最大深度和宽度。关于这些可能性及后续研究方向的更详细讨论，见 Hubin (2018)。最后，应研究变分推断在这些复杂非线性模型中的准确性。即使在线性模型中，Carbonetto 等人 (2012年) 也已表明，结果可能存在严重偏差。人们已提出多种减少变分推断偏差的方法。例如，可以采用更灵活的变分分布族，如引入辅助变量（Ranganath 等人，2016年；Salimans 等人，2015年）、归一化流（Louizos 和 Welling，2017年），或者采用刀切法消除偏差（Nowozin，2018）。这些方向留待未来研究。

本文提出的方法只能以隐式方式对架构进行宽度选择：对于 ReLU 等特定激活函数，该方法允许选择宽度；例如，若某个神经元的所有权重都被置零，即将给定神经元的所有 γ 设为 0，则该单元会从层中排除，从而缩小宽度。本文没有讨论神经网络的深度选择，但只要允许使用带跳跃连接的架构（即所有层都与响应以及网络的后续层相连），就可以采用类似的隐式方法选择深度。在这种情况下，可以将深度为 k 的特定层中的所有节点排除，使得只有较浅的层会影响响应。这样便可以推断深度不确定性。因此，研究深度选择是一个值得后续探讨的方向。此外，未来也可以像 Hubin 等人 (2021) 那样，通过指定针对性先验，更明确地处理深度和宽度选择。不过，与马尔可夫链蒙特卡罗相比，Hubin 等人 (2021) 的思路会给变分贝叶斯带来更多挑战。同时，正如 Hubin 等人 (2021) 所示，这类程序更有可能提供高度可解释的模型。此外，若要显式选择宽度和深度，还需要进一步研究模型先验。

最后但同样重要的是，我们想讨论一些在本文投稿和审稿期间发表的同期研究。首先，有两篇理论论文探讨了贝叶斯变分深度学习在一般情形和稀疏情形下的后验一致性（Bhattacharya 和 Maiti，2021；Chérief-Abdellatif，2020）。此外，Bai 等（2020）从理论上（在渐近框架下）论证了 Hubin（2018）以及 Hubin 和 Storvik（2019）所提出模型及其先验分布中先验纳入概率的选择。他们采用的变分推断技术与 Hubin 和 Storvik（2019）提出的方法几乎完全相同。Hubin 和 Storvik（2019）提出的方法近期还应用于遗传关联研究（Cheng 等，2022）。最后，Sun 等（2022）使用马尔可夫链蒙特卡罗对我们早期版本中 Hubin 和 Storvik（2019）提出的模型进行了推断。这些进展表明该领域发展迅速，也再次凸显了贝叶斯神经网络的方法学发展是多么贴近现实且与时俱进。

致谢

作者谨此感谢 Sean Murray（挪威计算中心）对文章语言提出的意见，并感谢 Pierre Lison 博士（挪威计算中心）就相关文献、潜在应用和技术工具所进行的深入讨论。我们还感谢 Petter Mostad 博士、查尔姆斯理工大学数学科学系及哥德堡大学的各位同仁对命题 2 提出的宝贵意见。我们也感谢在本文发表的各个阶段收到的审稿意见和编辑意见，这些意见富有建设性。

参考文献

Adya, M. 和 Collopy, F.（1998）。《神经网络在预测与预报方面的效果如何？一项综述与评估》。《预测学杂志》，17：481–495。1 Ahmed, A.、Aly, M.、Gonzalez, J.、Narayanamurthy, S. 和 Smola, A. J.（2012）。《潜变量模型中的可扩展推断》。载于第五届 ACM 网络搜索与数据挖掘国际会议论文集，第 123–132 页。ACM。3 Attias, H.（2000）。《图模型的变分贝叶斯框架》。载于《神经信息处理系统进展》，第 209–215 页。3 Bai, J.、Song, Q. 和 Cheng, G.（2020）。《具有理论保证的稀疏深度学习高效变分推断》。《神经信息处理系统进展》，33：466–476。2, 12, 23 Barbieri, M. M.、Berger, J. O.、George, E. I. 和 Ročková, V.（2021）。《中位概率模型与相关变量》。Bayesian Analysis，1–28。2 Barbieri, M. M.、Berger, J. O. 等（2004）。《最优预测模型选择》。《统计年鉴》，32(3)：870–897。2, 10 Bardenet, R.、Doucet, A. 和 Holmes, C.（2014）。《迈向可扩展的马尔可夫链蒙特卡罗：一种自适应子抽样方法》。载于《国际机器学习会议》，第 405–413 页。3 —（2017）。《关于大规模数据的马尔可夫链蒙特卡罗方法》。《机器学习研究杂志》，18(1)：1515–1557。3 Bhattacharya, S. 和 Maiti, T.（2021）。《变分贝叶斯神经网络的统计学基础》。《神经网络》，137：151–173。23 Bishop, C. M.（1997）。《贝叶斯神经网络》。《巴西计算机学会期刊》，4(1)。3 Blalock, D.、Gonzalez Ortiz, J. J.、Frankle, J. 和 Guttag, J.（2020）。《神经网络剪枝的现状如何？》《机器学习与系统会议论文集》，2：129–146。13 Blei, D. M.、Kucukelbir, A. 和 McAuliffe, J. D.（2017）。《变分推断：面向统计学家的综述》。《美国统计学会杂志》，112(518)：859–877。6 Blundell, C.、Cornebise, J.、Kavukcuoglu, K. 和 Wierstra, D.（2015）。《神经网络中的权重不确定性》。arXiv 预印本 arXiv:1505.05424。2, 5, 8, 9, 13 Bornkamp, B.、Ohlssen, D.、Magnusson, B. P. 和 Schmidli, H.（2017）。《亚组治疗效应估计的模型平均》。《药物统计学》，16(2)：133–142。4 Breiman, L.（2001）。《统计建模：两种文化（附评论及作者回应）》。《统计科学》，16(3)：199–231。3 Carbonetto, P.、Stephens, M. 等（2012）。《回归中贝叶斯变量选择的可扩展变分推断及其在遗传关联研究中的准确性》。Bayesian Analysis，7(1)：73–108。4, 6, 11, 23 Carvalho, C. M.、Polson, N. G. 和 Scott, J. G.（2009）。《通过马蹄先验处理稀疏性》。载于《人工智能与统计》，第 73–80 页。2 Cheng, W.、Ramachandran, S. 和 Crawford, L.（2022）。《利用贝叶斯神经网络进行遗传精细定位中变量选择的不确定性量化》。《iScience》，104553。23

Chérief-Abdellatif, B.-E. (2020)。《稀疏深度学习中变分推断的收敛速率》。载于《国际机器学习会议》，1831–1842。PMLR。23 Claeskens, G. 和 Hjort, N. L. (2008)。《模型选择与模型平均》。剑桥统计与概率数学系列。剑桥大学出版社。3 Claeskens, G.、Hjort, N. L. 等 (2008)。《模型选择与模型平均》。剑桥图书。2 Clyde, M. A.、Ghosh, J. 和 Littman, M. L. (2011)。《用于变量选择和模型平均的贝叶斯自适应抽样》。计算与图形统计学杂志，20(1)：80–101。4, 5 Dey, D. K.、Ghosh, S. K. 和 Mallick, B. K. (2000)。《广义线性模型：贝叶斯视角》。CRC出版社。5 Dobra, A.、Hans, C.、Jones, B.、Nevins, J. R.、Yao, G. 和 West, M. (2004)。《用于探索基因表达数据的稀疏图模型》。多元分析杂志，90(1)：196–212。21 Fahrmeir, L. 和 Lang, S. (2001)。《基于马尔可夫随机场先验的广义加性混合模型贝叶斯推断》。皇家统计学会杂志：C辑（应用统计），50(2)：201–220。21 Frommlet, F.、Ljubic, I.、Arnardóttir Helga, B. 和 Bogdan, M. (2012)。《使用以BIC修正形式作为适应度函数的模因算法进行QTL定位》。统计学在遗传学与分子生物学中的应用，11(4)：1–26。4, 5 Gal, Y. (2016)。《深度学习中的不确定性》。博士论文，剑桥大学。2, 3, 11 Gal, Y.、Hron, J. 和 Kendall, A. (2017)。《Concrete dropout》。载于《神经信息处理系统进展》，3581–3590。2, 8, 13 Geisser, S. 和 Eddy, W. F. (1979)。《一种模型选择的预测方法》。美国统计学会杂志，74(365)：153–160。3 George, E. I. 和 McCulloch, R. E. (1993)。《通过吉布斯抽样进行变量选择》。美国统计学会杂志，88(423)：881–889。4 George, E. I. 等 (2010)。《稀释先验：补偿模型空间冗余》。载于《借力：赋能应用的理论——献给 Lawrence D. Brown 的纪念文集》，158–165。数学统计学会。21 Ghosh, J. (2015)。《使用中位概率模型进行贝叶斯模型选择》。Wiley跨学科评论：计算统计学，7(3)：185–193。4 Ghosh, S.、Yao, J. 和 Doshi-Velez, F. (2018)。《使用马蹄先验对贝叶斯神经网络进行结构化变分学习》。载于《国际机器学习会议》，1744–1753。PMLR。2 — (2019)。《通过马蹄先验进行贝叶斯神经网络模型选择》。机器学习研究杂志，20(182)：1–46。2, 5 Goodfellow, I.、Bengio, Y. 和 Courville, A. (2016)。《深度学习》。麻省理工学院出版社。http://www.deeplearningbook.org。1 Graves, A. (2011)。《神经网络的实用变分推断》。载于《神经信息处理系统进展》，2348–2356。3, 4, 6, 13 Hansen, M. H. 和 Yu, B. (2001)。《模型选择与最小描述长度原则》。美国统计学会杂志，96(454)：746–774。3 Hastie, T.、Buja, A. 和 Tibshirani, R. (1995)。《惩罚判别分析》。统计学年刊，73–102。11 Heinze, G.、Wallisch, C. 和 Dunkler, D. (2018)。《变量选择——综述及对统计实践者的建议》。生物统计学杂志，60(3)：431–449。3

Hernández-Lobato, J. M., Hernández-Lobato, D., 和 Suárez, A.（2015）。“具有尖峰与 slab 先验的线性回归模型中的期望传播。” Machine Learning，99(3)：437–487。11 Hubin, A.（2018）。“复杂回归情境中的贝叶斯模型配置、选择与平均。”博士论文，奥斯陆大学。网址 https://www.duo.uio.no/handle/10852/65638 2, 4, 5, 23 Hubin, A. 和 Storvik, G.（2018）。“广义线性混合模型中的贝叶斯变量选择模式跳跃马尔可夫链蒙特卡罗。” Computational Statistics & Data Analysis。4 —（2019）。“贝叶斯神经网络中的模型与参数不确定性结合。”arXiv 预印本 arXiv:1903.07594。2, 5, 12, 23 Hubin, A., Storvik, G., 和 Frommlet, F.（2021）。“灵活的贝叶斯非线性模型配置。” Journal of Artificial Intelligence Research，72：901–942。3, 4, 5, 23 Hubin, A., Storvik, G., Frommlet, F. 等（2020）。“贝叶斯逻辑回归的一种新颖算法方法（附讨论）。” Bayesian Analysis，15(1)：263–333。4 Humphreys, K. 和 Titterington, D.（2000）。“简单混合模型的近似贝叶斯推断。”载于 Computational Statistics 2000 会议录，331–336。3 Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., 和 Saul, L. K.（1999）。“图模型变分方法简介。”Machine learning，37(2)：183–233。3 Jylänki, P., Nummenmaa, A., 和 Vehtari, A.（2014）。“具有促进稀疏性的先验分布的神经网络期望传播。”Journal of Machine Learning Research，15：1849–1901。网址 http://jmlr.org/papers/v15/jylanki14a.html 2 Kanter, J. M. 和 Veeramachaneni, K.（2015）。“深度特征合成：迈向数据科学工作的自动化。”载于 Data Science and Advanced Analytics (DSAA), 2015。36678 2015。IEEE 国际会议，1–10。IEEE。1 Kingma, D. P. 和 Ba, J.（2014）。“Adam：一种随机优化方法。”arXiv 预印本 arXiv:1412.6980。11 Kingma, D. P., Salimans, T., 和 Welling, M.（2015）。“变分丢弃法与局部重参数化技巧。”载于神经信息处理系统进展，2575–2583。8 Korattikara, A., Chen, Y., 和 Welling, M.（2014）。“MCMC领域的节俭：削减 Metropolis-Hastings 预算。”载于国际机器学习会议，181–189。3 Kuo, L. 和 Mallick, B.（1998）。“回归模型的变量选择。”Sankhy ¯a：The Indian Journal of Statistics，B辑，65–81。3 LeCun, Y., Cortes, C., 和 Burges, C.（1998）。“MNIST手写数字数据库，1998。”网址 http://www. research. att. com/ ˜yann/ocr/mnist。11 Li, Y. 和 Clyde, M. A.（2018）。“广义线性模型中的 g 先验混合。”美国统计学会杂志，113(524)：1828–1845。5 Liu, S., Mingas, G., 和 Bouganis, C.-S.（2015）。“自定义精度机制下的精确MCMC加速器。”载于 Field Programmable Technology (FPT), 2015 国际会议，120–127。IEEE。3 Logsdon, B. A., Hoffman, G. E., 和 Mezey, J. G.（2010）。“用于快速准确进行多位点全基因组关联分析的变分贝叶斯算法。”BMC bioinformatics，11(1)：58。4, 6

Louizos, C., Ullrich, K. 和 Welling, M.（2017）。“深度学习的贝叶斯压缩。” 收录于《神经信息处理系统进展》，3288–3298。2，4，5，13 Louizos, C. 和 Welling, M.（2017）。“用于变分贝叶斯神经网络的乘性归一化流。” 收录于第34届国际机器学习会议论文集（第70卷），2218–2227。JMLR.org。19，23 MacKay, D. J.（1995）。“贝叶斯神经网络与密度网络。”《核仪器与方法物理研究A辑：加速器、谱仪、探测器及相关设备》，354(1)：73–80。3 —（1997）。“隐马尔可夫模型的集成学习。”未发表手稿。网址 https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.52.9627&rep=rep1&type=pdf 3 Maclaurin, D. 和 Adams, R. P.（2014）。“萤火虫蒙特卡洛：使用数据子集进行精确的马尔可夫链蒙特卡洛。” 收录于 UAI，543–552。3 Mitchell, T. J. 和 Beauchamp, J. J.（1988）。“线性回归中的贝叶斯变量选择。”《美国统计学会杂志》，83(404)：1023–1032。4 Molchanov, D., Ashukha, A. 和 Vetrov, D.（2017）。“变分丢弃法使深度神经网络稀疏化。” 收录于第34届国际机器学习会议论文集（第70卷），2498–2507。JMLR.org。2 Neal, R. M.（1992）。“通过混合蒙特卡洛方法对反向传播网络进行贝叶斯训练。”技术报告，Citeseer。3 Neklyudov, K., Molchanov, D., Ashukha, A. 和 Vetrov, D.（2018）。“方差网络：当期望不如所愿。” 收录于国际学习表征会议。2 Neklyudov, K., Molchanov, D., Ashukha, A. 和 Vetrov, D. P.（2017）。“通过对数正态乘性噪声进行结构化贝叶斯剪枝。” 收录于《神经信息处理系统进展》，6775–6784。2 Nitarshan, R.（2018）。“神经网络中的权重不确定性。” https://www.nitarshan.com/bayes-by-backprop/。17 Nowozin, S.（2018）。“证据近似的去偏：关于重要性加权自编码器与刀切法变分推断。” 收录于国际学习表征会议。23 Palvanov, A. 和 Im Cho, Y.（2018）。“实时环境中 MNIST 深度学习算法的比较。”《模糊逻辑与智能系统国际期刊》，18(2)：126–134。14 Polson, N. 和 Rockova, V.（2018）。“稀疏深度学习的后验集中。”arXiv 预印本 arXiv:1803.09138。2，4，5 Posch, K., Steinbrener, J. 和 Pilz, J.（2019）。“通过变分推断测量深度神经网络中的模型不确定性。”arXiv 预印本 arXiv:1902.10189。13，21 Quiroz, M., Kohn, R., Villani, M. 和 Tran, M.-N.（2014）。“通过高效的数据子抽样加速马尔可夫链蒙特卡洛。”arXiv 预印本 arXiv:1404.4178。3 Quiroz, M., Villani, M. 和 Kohn, R.（2016）。“精确子抽样马尔可夫链蒙特卡洛。”arXiv 预印本 arXiv:1603.08232。3 Ranganath, R., Tran, D. 和 Blei, D.（2016）。“层次变分模型。” 收录于国际机器学习会议，324–333。23 Razi, M. A. 和 Athappilly, K.（2005）。“神经网络（NN）、非线性回归以及分类回归树（CART）模型的比较预测分析。”《专家系统及其应用》，29(1)：65–74。1

Refenes, A. N., Zapranis, A., 和 Francis, G. (1994). “使用神经网络进行股票表现建模：与回归模型的比较研究。”《神经网络》，7(2): 375–388。1 Ripley, B. D. (2007).《模式识别与神经网络》。剑桥大学出版社。13 Salimans, T., Kingma, D., 和 Welling, M. (2015). “马尔可夫链蒙特卡罗与变分推断：弥合差距。”载于《国际机器学习会议》，1218–1226。23 Sargent, D. J. (2001). “人工神经网络与其他统计方法的比较。”Cancer，91(S8): 1636–1642。1 Smith, T. E. 和 LeSage, J. P. (2004). “具有空间依赖关系的贝叶斯概率单位模型。”载于《空间与时空计量经济学》，127–160。Emerald Group Publishing Limited。21 Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., 和 Salakhutdinov, R. (2014). “丢弃法：防止神经网络过拟合的简单方法。”《机器学习研究杂志》，15(1): 1929–1958。2, 7 Steel, M. F. (2020). “模型平均及其在经济学中的应用。”《经济文献杂志》，58(3): 644–719。3, 4 Sun, Y., Song, Q., 和 Liang, F. (2022). “利用尖峰与平板先验学习稀疏深度神经网络。”《统计与概率快报》，180: 109246。23 Wasserman, L. (2000). “贝叶斯模型选择与模型平均。”《数学心理学杂志》，44(1): 92–107。10 Welling, M. 和 Teh, Y. W. (2011). “通过随机梯度朗之万动力学进行贝叶斯学习。”载于《第28届国际机器学习大会论文集》(ICML-11)，681–688。3 Xiao, H., Rasul, K., 和 Vollgraf, R. (2017). “Fashion-MNIST：用于机器学习算法基准测试的新型图像数据集。”arXiv 预印本 arXiv:1708.07747 . 11

附录 A：LBBNN-GP-LFMVN 的部分调优参数与结果

	A_β, A_ρ	A_ξ	A_ω	A_Σ	A_{a_ψ},A_{b_ψ}	A_{a_β},A_{b_β}
预训练	0.00010	0.01000	-	0.01000	0.00100	0.00001
训练	0.00010	0.00010	-	0.00010	0.00000	0.00000
后训练	0.00010	0.00000	-	0.00000	0.00000	0.00000

表 7： LBBNN-GP-LFMVN 优化算法步长的 $\boldsymbol{A}$ 矩阵对角元素规格。

表 8： MNIST 数据的性能指标，补充表 3，展示在变分推断部分使用低秩因子多元正态分布的结果。得出这些结果时未进行后训练。更多详情请参见正文中的表 2 和表 3 的标题。

预测 γ	β	模型先验-方法	R	所有类别 Acc	0.95 阈值 Acc	0.95 阈值 类别数	密度级别	Epo. time
SIM	SIM	LBBNN-GP-LFMVN	1	0.959 (0.956,0.960)	-	-	0.450	13.009
SIM	SIM	LBBNN-GP-LFMVN	10	0.979（0.978,0.980）	1.000	7760	1.000	13.009
ALL	EXP	LBBNN-GP-LFMVN	1	0.976 (0.976,0.978)	-	-	1.000	13.009
MED	SIM	LBBNN-GP-LFMVN	1	0.959 (0.956,0.960)	-	-	0.449	13.009
MED	SIM	LBBNN-GP-LFMVN	10	0.979（0.978,0.980）	1.000	7764	0.449	13.009
MED	EXP	LBBNN-GP-LFMVN	1	0.975 (0.974,0.977)	-	-	0.449	13.009

表 9：FMNIST 数据的性能指标，补充至表 4 ，在变分推断部分使用低秩因子多元正态分布。获得这些结果时未应用后训练。详情请参阅正文中的表 2 和表 3 的标题。

预测 γ	β	模型先验-方法	R	所有类别 Acc	0.95 阈值 Acc	0.95 阈值 类别数	密度级别	Epo. time
SIM	SIM	LBBNN-GP-LFMVN	1	0.849（0.845，0.853）	-	-	0.456	12.949
SIM	SIM	LBBNN-GP-LFMVN	10	0.876 (0.874,0.880)	0.996	4485	1.000	12.949
ALL	EXP	LBBNN-GP-LFMVN	1	0.864（0.862,0.867）	-	-	1.000	12.949
MED	SIM	LBBNN-GP-LFMVN	1	0.849（0.844,0.852）	-	-	0.455	12.949
MED	SIM	LBBNN-GP-LFMVN	10	0.877 (0.875,0.878)	0.996	4486	0.455	12.949
MED	EXP	LBBNN-GP-LFMVN	1	0.862（0.858，0.864）	-	-	0.455	12.949

表10： 在变分推断部分使用低秩因子多元正态分布，为表5 添加 PHONEME 数据的性能指标。获得这些结果时未进行后训练。详情请参见正文中的表 2 和表 3 的标题。

预测		模型先验-		所有类别	0.95 阈值		密度	Epo.
γ	β	方法	R	Acc	Acc	类别数	级别	time
SIM	SIM	LBBNN-GP-LFMVN	1	0.918（0.906，0.928）	-	-	0.497	0.472
SIM	SIM	LBBNN-GP-LFMVN	10	0.929 (0.926,0.934)	0.994	663	1.000	0.472
ALL	EXP	LBBNN-GP-LFMVN	1	0.921 (0.915,0.931)	-	-	1.000	0.472
MED	SIM	LBBNN-GP-LFMVN	1	0.918 (0.909,0.930)	-	-	0.497	0.472
MED	SIM	LBBNN-GP-LFMVN	10	0.929（0.925，0.933）	0.995	664	0.497	0.472
MED	EXP	LBBNN-GP-LFMVN	1	0.917 (0.912,0.925)	-	-	0.497	0.472

附录 B：不同剪枝程度下频率学派神经网络的结果

在本实验中，我们纳入了频率学派神经网络的标准基于幅度的剪枝结果。该网络与论文正文中用于 MNIST、FMNIST 和 PHONEMNE 数据集的贝叶斯神经网络采用相同配置。在表 S-5、S-6 和 S-7 中，我们展示了：对于所涉及的三个数据集，在全部稀疏度水平（对应于通过贝叶斯方法获得的稀疏度水平）下，就预测准确率而言，所有贝叶斯方法均优于频率学派对应方法。然而，全连接的稠密 FNN 的表现与贝叶斯版本相当。此处的稀疏性是通过移除幅度最低的相应比例的权重/神经元获得的。

表 11： 不同剪枝程度下频率学派神经网络在 MNIST 数据上的性能指标。

密度水平	A相应的神经元剪枝	Acc. 权重剪枝	每轮时间
1.000	98.110 (98.040,98.350)	98.110（98.040，98.350）	1.360
0.500	98.030（97.720，98.080）	87.170 (78.260,90.730)	1.360
0.226	96.600 (95.570,97.150)	36.915（32.600，46.910）	1.360
0.194	96.130（94.680，96.670）	32.320（27.320，38.150）	1.360
0.180	95.710 (93.290,96.730)	31.685（25.010，35.840）	1.360
0.163	95.070（91.040，96.170）	29.760（23.810，33.340）	1.360
0.090	80.050 (75.060,90.120)	18.235 (14.500,21.520)	1.360
0.079	77.730 (68.910,86.090)	19.055 (15.270,25.990)	1.360

表 12：FMNIST 数据上不同剪枝程度下频率学派神经网络的性能指标。

密度水平	神经元剪枝准确率	权重剪枝准确率	每轮时间
1.000	89.470（86.240，89.790）	89.470（86.240，89.790）	1.260
0.500	85.150 (80.040,87.180)	29.865 (21.370,41.080)	1.260
0.302	67.610（53.110，75.760）	17.175（14.020，20.660）	1.260
0.156	41.595 (37.520,45.510)	11.460（9.370，20.720）	1.260
0.129	39.040 (30.090,44.980)	10.515 (9.300,18.490)	1.260
0.120	39.775 (26.140,44.480)	09.980 (8.130,18.610)	1.260
0.108	38.750（23.890，43.660）	10.325（8.080，20.110）	1.260
0.094	36.340（23.640，41.050）	09.790（8.040，25.000）	1.260

表 13： 频率学派神经网络的性能指标在不同剪枝程度下对 PHONEMNE 数据的表现

密度水平	准确率 神经元剪枝	准确率 权重剪枝	每轮时间
1.000	92.400（92.1,92.9）	92.500 (92.200,92.700)	0.029
0.600	92.250（91.8，93.2）	88.050 (86.100,90.500)	0.029
0.509	92.100 (91.5,92.7)	82.650 (78.100,85.400)	0.029
0.457	92.100（91.6，92.7）	80.950（75.700，86.500）	0.029
0.371	91.850 (90.4,92.6)	76.050（68.300，82.200）	0.029
0.307	91.700（90.7，92.3）	71.950（66.500，80.200）	0.029
0.255	91.000（90.4，92.8）	67.900（58.600，80.200）	0.029
0.225	90.700（90.2，92.6）	64.950（51.800，77.300）	0.029

附录 C：基于后训练的结果

表 14： 文中所提出的 BNN 贝叶斯方法在 MNIST 数据上的性能指标。此处报告了后训练后的结果。更多详情请参见表 2 和表 3 的标题。

				MNIST			
				所有类别	0.95 阈值		密度
γ	β	方法	R	Acc	Acc	类别数	级别
SIM	SIM	LBBNN-GP-MF	1	0.967（0.966，0.969）	-	-	0.090
SIM	SIM	LBBNN-GP-MF	10	0.980 (0.979,0.982)	0.999	8346	1.000
ALL	EXP	LBBNN-GP-MF	1	0.982（0.980，0.983）	-	-	1.000
MED	SIM	LBBNN-GP-MF	1	0.969（0.966，0.972）	-	-	0.079
MED	SIM	LBBNN-GP-MF	10	0.980 (0.979,0.982)	0.999	8472	0.079
MED	EXP	LBBNN-GP-MF	1	0.981（0.980，0.984）	-	-	0.079
SIM	SIM	LBBNN-GP-MVN	1	0.967（0.965，0.970）	-	-	0.180
SIM	SIM	LBBNN-GP-MVN	10	0.979（0.977，0.981）	1.000	7994	1.000
ALL	EXP	LBBNN-GP-MVN	1	0.980（0.979，0.981）	-	-	1.000
MED	SIM	LBBNN-GP-MVN	1	0.973（0.971，0.977）	-	-	0.163
MED	SIM	LBBNN-GP-MVN	10	0.978 (0.977, 0.979)	1.000	8107	0.163
MED	EXP	LBBNN-GP-MVN	1	0.976 (0.974,0.977)	-	-	0.163
SIM	SIM	LBBNN-GP-LFMVN	1	0.971（0.969，0.973）	-	-	0.450
SIM	SIM	LBBNN-GP-LFMVN	10	0.978（0.976，0.980）	0.999	8366	1.000
ALL	EXP	LBBNN-GP-LFMVN	1	0.979 (0.978,0.980)	-	-	1.000
MED	SIM	LBBNN-GP-LFMVN	1	0.973（0.971，0.977）	-	-	0.449
MED	SIM	LBBNN-GP-LFMVN	10	0.978 (0.977, 0.979)	0.999	8645	0.449
MED	EXP	LBBNN-GP-LFMVN	1	0.978（0.976，0.979）	-	-	0.449
SIM	SIM	BNN-GP-CMF	1	0.982 (0.894,0.984)	-	-	0.226
SIM	SIM	BNN-GP-CMF	10	0.984 (0.896,0.986)	0.995	9586	1.000
ALL	EXP	BNN-GP-CMF	1	0.983（0.894，0.984）	-	-	1.000
PRN	SIM	BNN-HP-MF	1	0.967 (0.965,0.968)	-	-	0.194
PRN	SIM	BNN-HP-MF	10	0.982（0.981，0.983）	1.000	0007	0.194
PRN	EXP	BNN-HP-MF	1	0.966（0.964，0.969）	-	-	0.194

表15：本文提出的 BNN 贝叶斯方法在 FMNIST 数据上的性能指标。此处报告后训练后的结果。更多详情请参见表 2 和表 3 的标题。

				FMNIST			
				所有类别	0.95 阈值		密度
γ	β	方法	R	Acc	Acc	类别数	级别
SIM	SIM	LBBNN-GP-MF	1	0.863 (0.862,0.869)	-	-	0.120
SIM	SIM	LBBNN-GP-MF	10	0.884 (0.881,0.886)	0.995	4932	1.000
ALL	EXP	LBBNN-GP-MF	1	0.882（0.879，0.887）	-	-	1.000
MED	SIM	LBBNN-GP-MF	1	0.866 (0.865,0.869)	-	-	0.108
MED	SIM	LBBNN-GP-MF	10	0.885 (0.882,0.887)	0.995	4994	0.108
MED	EXP	LBBNN-GP-MF	1	0.882 (0.878,0.886)	-	-	0.108
SIM	SIM	LBBNN-GP-MVN	1	0.859 (0.857,0.862)	-	-	0.156
SIM	SIM	LBBNN-GP-MVN	10	0.880 (0.874,0.881)	0.996	4615	1.000
ALL	EXP	LBBNN-GP-MVN	1	0.876 (0.872,0.877)	-	-	1.000
MED	SIM	LBBNN-GP-MVN	1	0.863（0.860, 0.865）	-	-	0.129
MED	SIM	LBBNN-GP-MVN	10	0.878（0.874，0.880）	0.995	4801	0.129
MED	EXP	LBBNN-GP-MVN	1	0.873（0.870，0.875）	-	-	0.129
SIM	SIM	LBBNN-GP-LFMVN	1	0.847（0.844，0.850）	-	-	0.456
SIM	SIM	LBBNN-GP-LFMVN	10	0.875（0.873，0.878）	0.996	4431	1.000
ALL	EXP	LBBNN-GP-LFMVN	1	0.866 (0.859,0.868)	-	-	1.000
MED	SIM	LBBNN-GP-LFMVN	1	0.846 (0.844,0.849)	-	-	0.455
MED	SIM	LBBNN-GP-LFMVN	10	0.876（0.873，0.879）	0.996	4426	0.455
MED	EXP	LBBNN-GP-LFMVN	1	0.864（0.860，0.865）	-	-	0.455
SIM	SIM	BNN-GP-CMF	1	0.897 (0.820,0.899)	-	-	0.094
SIM	SIM	BNN-GP-CMF	10	0.897（0.823，0.902）	0.943	8826	1.000
ALL	EXP	BNN-GP-CMF	1	0.896（0.820，0.901）	-	-	1.000
PRN	SIM	BNN-HP-MF	1	0.867（0.864，0.871）	-	-	0.302
PRN	SIM	BNN-HP-MF	10	0.888（0.887，0.890）	1.000	0147	0.302
PRN	EXP	BNN-HP-MF	1	0.868（0.864，0.869）	-	-	0.302

表 16： 文中提出的 BNN 贝叶斯方法在 PHONEME 数据上的性能指标。此处报告后训练后的结果。更多详情请参见表 2 和表 3 的说明。

				FMNIST			
				所有类别	0.95 阈值		密度
γ	β	方法	R	Acc	Acc	类别数	级别
SIM	SIM	LBBNN-GP-MF	1	0.917 (0.895,0.928)	-	-	0.120
SIM	SIM	LBBNN-GP-MF	10	0.927（0.921，0.931）	0.991	703	1.000
ALL	EXP	LBBNN-GP-MF	1	0.924（0.923，0.929）	-	-	1.000
MED	SIM	LBBNN-GP-MF	1	0.924（0.910，0.930）	-	-	0.108
MED	SIM	LBBNN-GP-MF	10	0.924 (0.911，0.932)	0.981	772	0.108
MED	EXP	LBBNN-GP-MF	1	0.925（0.909，0.931）	-	-	0.108
SIM	SIM	LBBNN-GP-MVN	1	0.922（0.914，0.928）	-	-	0.255
SIM	SIM	LBBNN-GP-MVN	10	0.931（0.926，0.935）	0.993	682	1.000
ALL	EXP	LBBNN-GP-MVN	1	0.927 (0.917,0.932)	-	-	1.000
MED	SIM	LBBNN-GP-MVN	1	0.923 (0.919,0.931)	-	-	0.225
MED	SIM	LBBNN-GP-MVN	10	0.928 (0.924, 0.933)	0.991	691	0.225
MED	EXP	LBBNN-GP-MVN	1	0.924 (0.915,0.930)	-	-	0.225
SIM	SIM	LBBNN-GP-LFMVN	1	0.919 (0.910,0.926)	-	-	0.497
SIM	SIM	LBBNN-GP-LFMVN	10	0.929（0.925，0.931）	0.994	678	1.000
ALL	EXP	LBBNN-GP-LFMVN	1	0.921 (0.912, 0.933)	-	-	1.000
MED	SIM	LBBNN-GP-LFMVN	1	0.917 (0.895,0.926)	-	-	0.497
MED	SIM	LBBNN-GP-LFMVN	10	0.928 (0.923,0.932)	0.994	676	0.497
MED	EXP	LBBNN-GP-LFMVN	1	0.917 (0.909,0.926)	-	-	0.497
SIM	SIM	BNN-GP-CMF	1	0.882 (0.716,0.904)	-	-	0.509
SIM	SIM	BNN-GP-CMF	10	0.921（0.918，0.930）	0.960	189	1.000
ALL	EXP	BNN-GP-CMF	1	0.877（0.697,0.909）	-	-	1.000
PRN	SIM	BNN-HP-MF	1	0.913 (0.909,0.926)	-	-	0.457
PRN	SIM	BNN-HP-MF	10	0.917 (0.916,0.922)	0.936	037	0.457
PRN	EXP	BNN-HP-MF	1	0.917 (0.914,0.924)	-	-	0.457

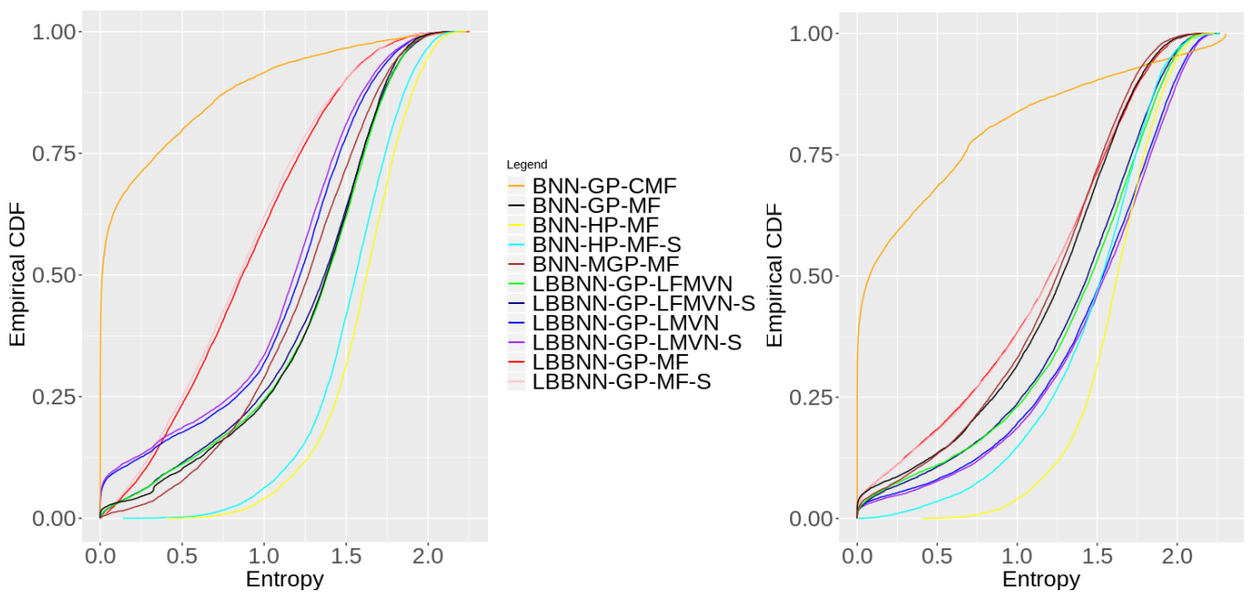

图 6：在 MNIST 上训练并应用于 FMNIST（左图），以及反向应用（右图）的边缘后验预测分布的熵的经验累积分布函数，结果对应于后训练后的仿真 $s = 10$ 。后缀 S 表示通过适当方法进行稀疏化的模型。

附录 D：先验参数固定时 LBBNN 的结果

本节补充了采用基础 LBBNN 获得的 MNIST 和 FMNIST 数据集结果。该模型对权重使用高斯先验，对权重纳入情况使用伯努利先验，不假设其他超先验。此处的先验纳入概率对应 BIC 惩罚项，即 $\psi^{(l)} = \exp(-2\log\ n)$，slav 分量的先验方差设为 1。此外，此处使用基础平均场变分分布。

表 17： 本文提出的 BNN 贝叶斯方法在 MNIST 数据上的性能指标。此处报告的是未经后训练的结果。详情请参见表 2 和表 3 的说明。

				MNIST			
				所有类别	0.95 阈值		
γ	β	方法	R	Acc	Acc	类别数	密度级别
SIM	SIM	LBBNN-GP-MF	1	0.958（0.954，0.960）	-	-	0.056
SIM	SIM	LBBNN-GP-MF	10	0.967 (0.966,0.971)	0.999	7064	0.084
ALL	EXP	LBBNN-GP-MF	1	0.969 (0.967,0.970)	-	-	1.000
MED	SIM	LBBNN-GP-MF	1	0.961 (0.957,0.964)	-	-	0.051
MED	SIM	LBBNN-GP-MF	10	0.964 (0.962,0.967)	0.998	7441	0.051
MED	EXP	LBBNN-GP-MF	1	0.965 (0.963,0.968)	-	-	0.051

表18：本文针对 BNN 的贝叶斯方法所提出的 FMNIST 数据的性能指标。此处报告的是未经后训练的结果。更多详细信息请参见表 2 和表 3 的图注。

				FMNIST			
				所有类别	0.95 阈值		
γ	β	方法	R	Acc	Acc	类别数	密度级别
SIM	SIM	LBBNN-GP-MF	1	0.854 (0.850,0.858)	-	-	0.066
SIM	SIM	LBBNN-GP-MF	10	0.867（0.863，0.870）	0.996	4097	0.083
ALL	EXP	LBBNN-GP-MF	1	0.866 (0.864,0.874)	-	-	1.000
MED	SIM	LBBNN-GP-MF	1	0.858（0.854，0.865）	-	-	0.065
MED	SIM	LBBNN-GP-MF	10	0.863 (0.859,0.869)	0.993	4347	0.065
MED	EXP	LBBNN-GP-MF	1	0.863（0.859，0.870）	-	-	0.065

表19： 在 MNIST 和 FMNIST 数据上进行 10 次重复实验所得模型的平均（每层）边际纳入概率的中位数和标准差（定义见正文）。

	MNIST 数据		**FMNIST 数据**	
层	Med.	SD.	Med.	SD.
$\rho(\gamma^{(1)}\|\mathcal{D})$	0.0520	0.0005	0.0665	0.0004
$\rho(\gamma^{(2)}\|\mathcal{D})$	0.0598	0.0003	0.0613	0.0005
$\rho(\gamma^{(3)}\|\mathcal{D})$	0.2217	0.0064	0.2013	0.0051

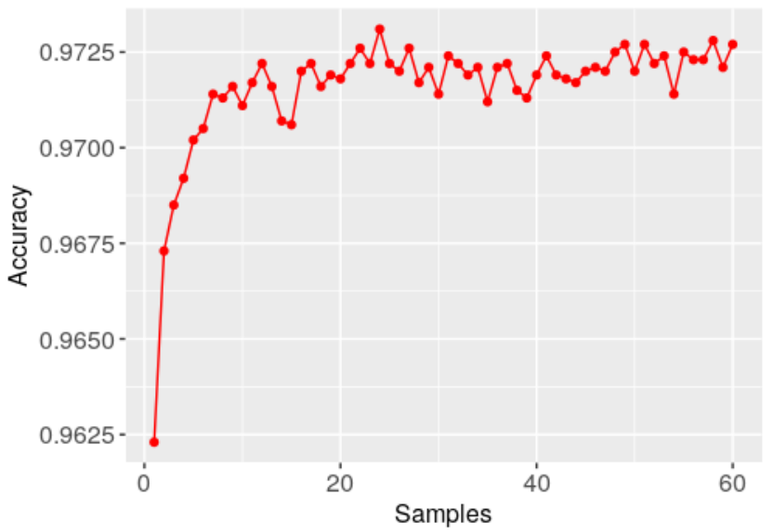

图 7：MNIST 数据上仿真 $i = 10$ 的预测准确率与模型及参数联合后验分布的样本数量之间的关系。

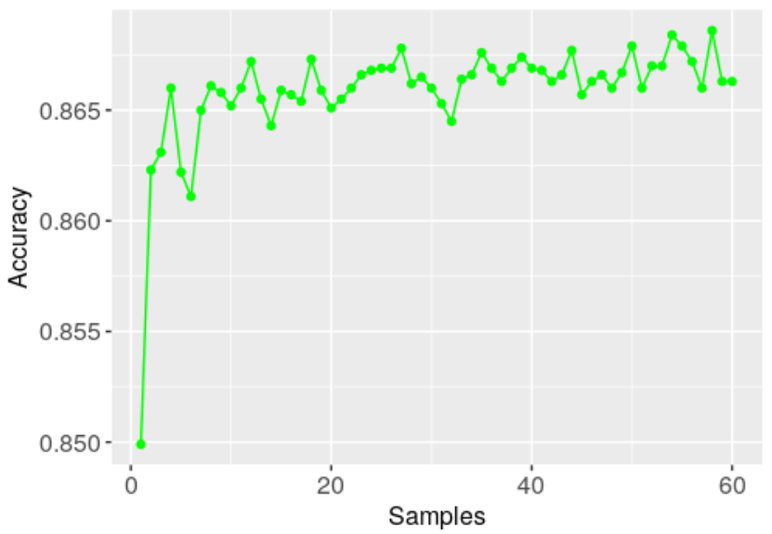

图 8： FMNIST 数据的联合后验分布样本数量与预测准确率之间的关系。

附录 E：不同层宽度下的结果

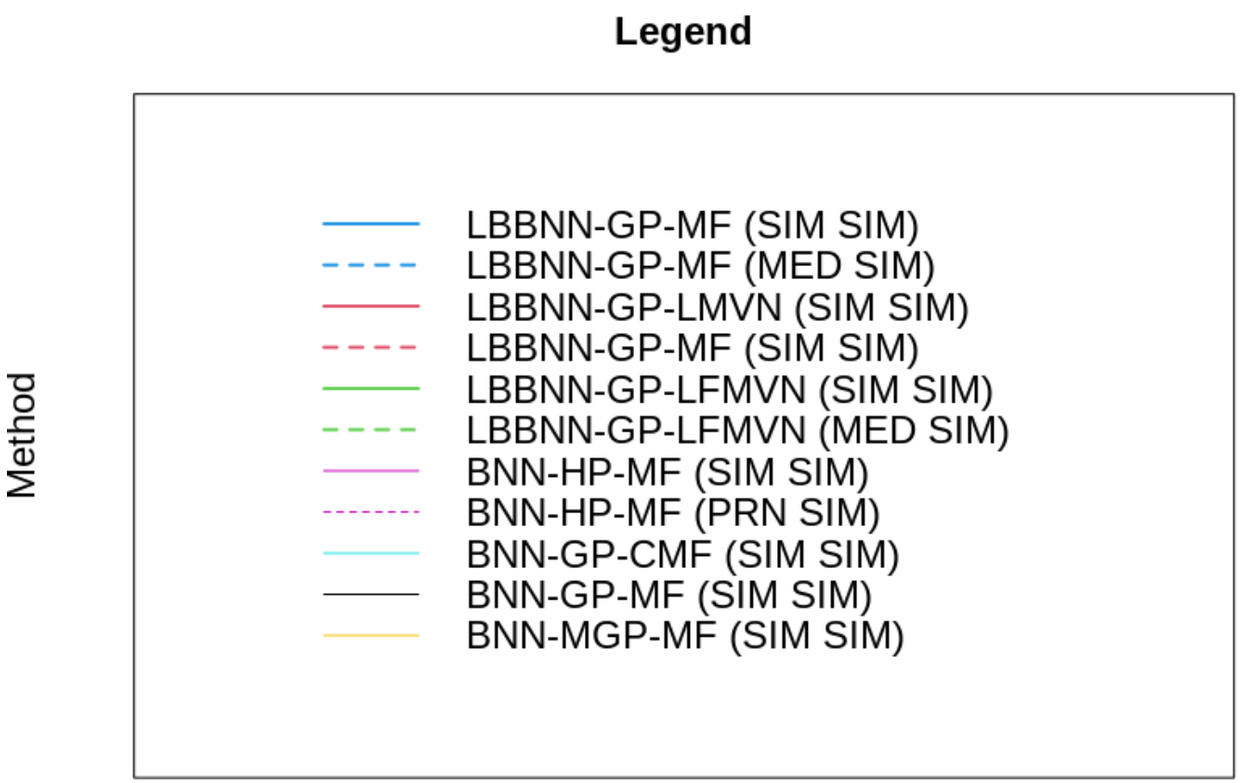

图 9： 图中所用线条的图例 10-12PHONEMNE 数据上宽度因子变化实验。

在本实验中，我们采用了主论文第 4 节 PHONEMNE 数据研究所使用的设置，但将所有层的宽度设为 40B-60B-60B，并将 B 从 1 改为 10。结果汇总于图 10-12。正如预期，所有稀疏性诱导方法的结果都表明，随着宽度增加，稀疏度水平也随之提高。总体而言，预测结果从 $\mathrm{B} = 1$ 到 $\mathrm{B} = 4$ 增加，并在 $\mathrm{B} = 4$处趋于稳定。最后，对于每个固定的 B，我们都得出了与主论文中 PHONEMNE 数据集所报告结果相同的总体结论：对于模型的所有宽度配置，LBBNN 都在预测准确率、不确定性感知准确率和稀疏性之间实现了良好且稳定的权衡。

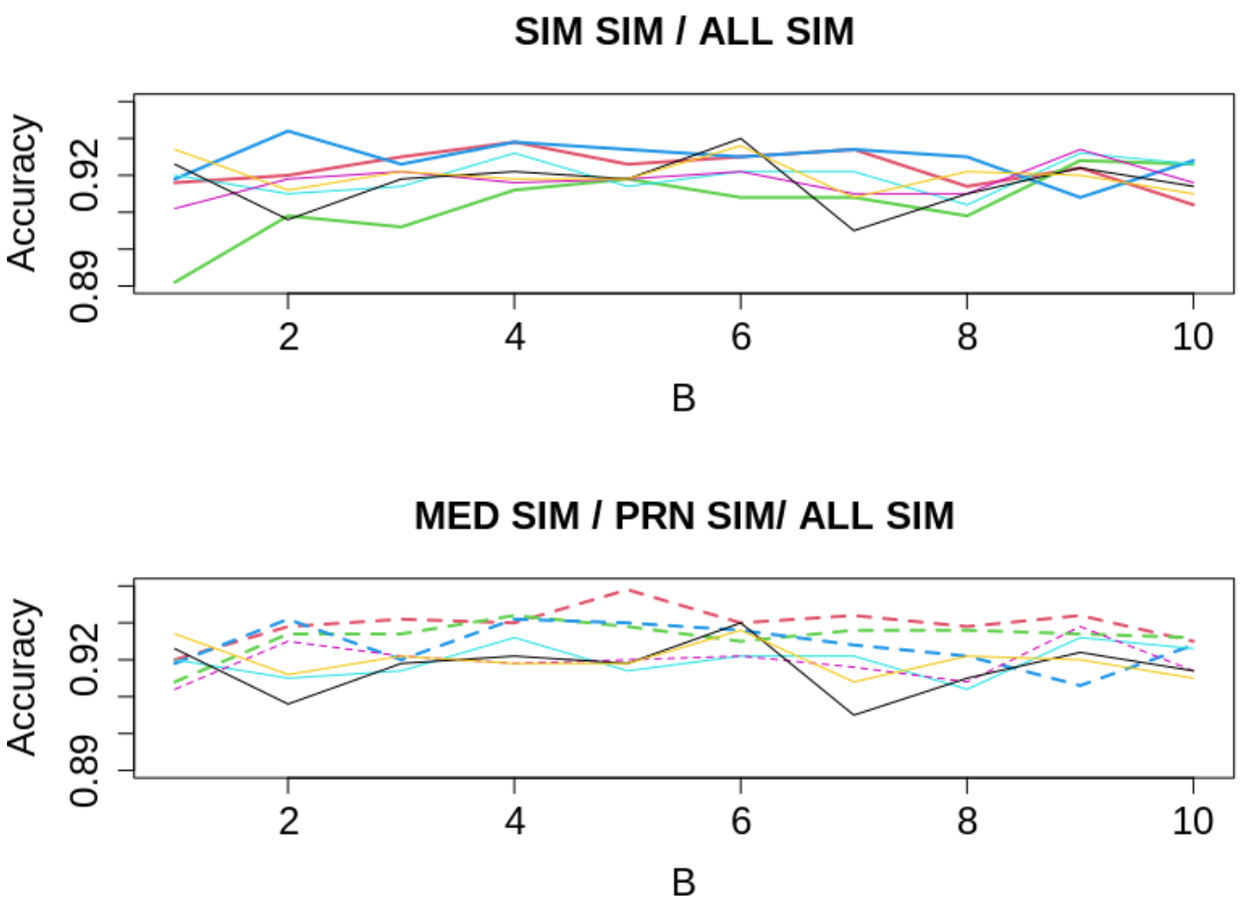

图10：PHONEMNE 数据的预测准确率与宽度因子的关系：上图：（SIM，SIM）对应 LBBNN-GP-MF、LBBNN-GP-MVN、LBBNN-GP-LF MVN、BNN-GP-CMF、BNN-HP-MF；（全部，SIM）对应 BNN-GP-MF、BNN-MGP-MF。下图：（MED，SIM）对应 LBBNN-GP-MF、LBBNN-GP-MVN、LBBNN-GP-LFMVN；（SIM SIM）对应 BNN-GP-CMF；（PRN SIM）对应 BNN-HP-MF；（全部，SIM）对应 BNN-GP-MF、BNN-MGP-MF。

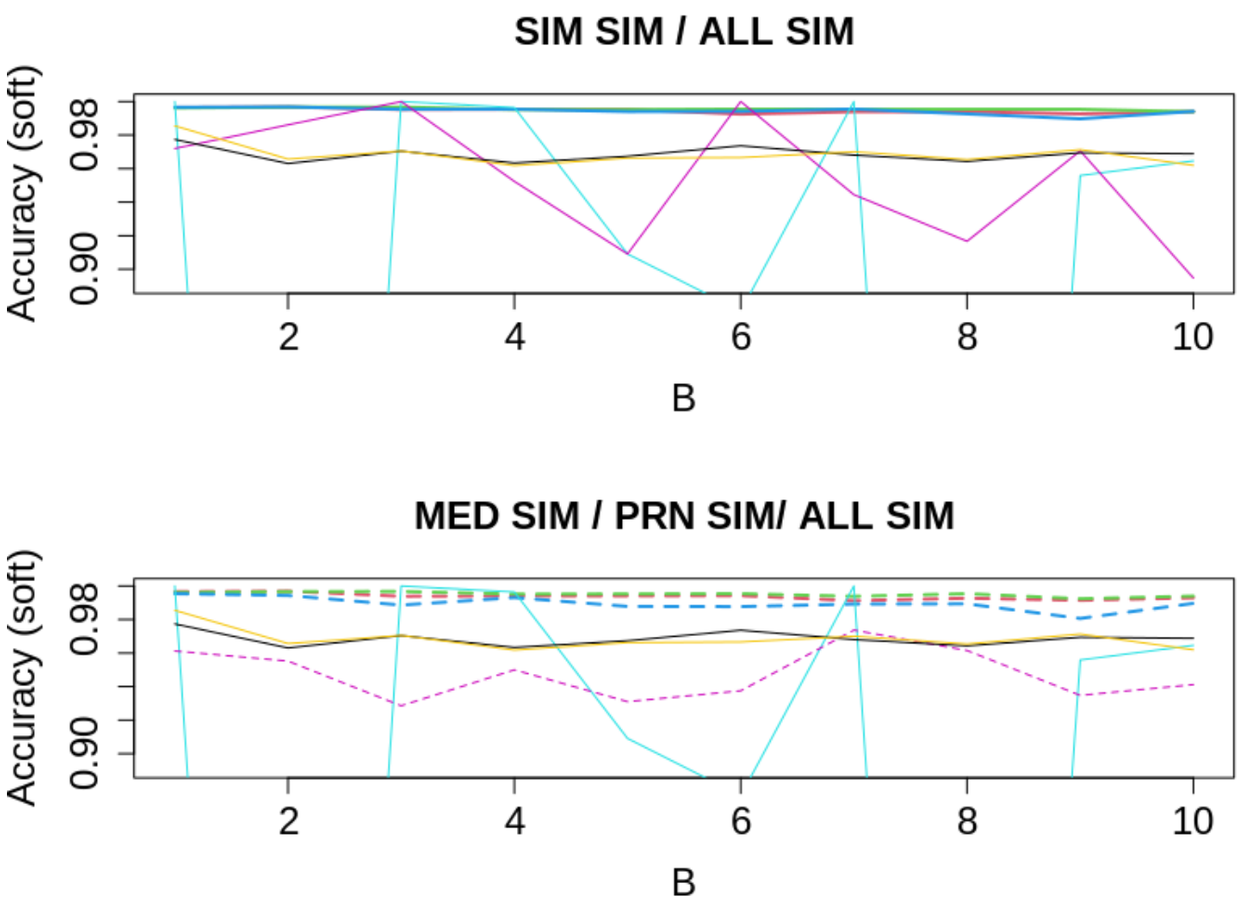

图 11：PHONEMNE 数据的不确定性感知预测准确率（后验模型平均概率阈值为 0.95）与宽度因子的关系：上图中，LBBNN-GP-MF、LBBNN-GP-MVN、LBBNN-GP-LFMVN、BNN-GP-CMF 和 BNN-HP-MF 对应 (SIM, SIM)；BNN-GP-MF 和 BNN-MGP-MF 对应 (全部, SIM)。下图中，LBBNN-GP-MF、LBBNN-GP-MVN 和 LBBNN-GP-LFMVN 对应 (MED, SIM)；BNN-GP-CMF 对应 (SIM SIM)；BNN-HP-MF 对应 (PRN SIM)；BNN-GP-MF 和 BNN-MGP-MF 对应 (全部, SIM)。

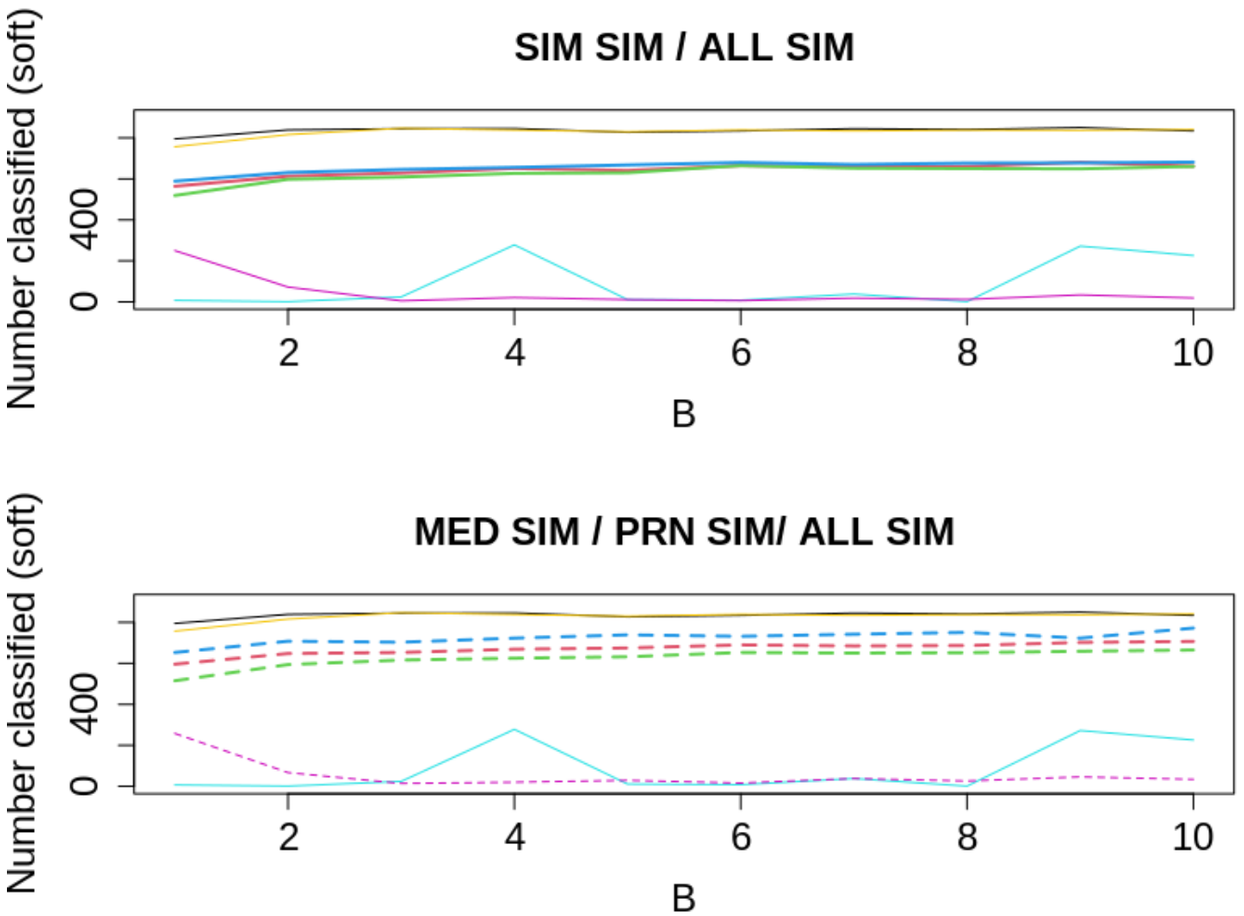

图 12： PHONEMNE 数据中不确定性感知预测的数量（后验模型平均概率阈值为 0.95）与宽度因子的关系：更多详情请参见图 11。

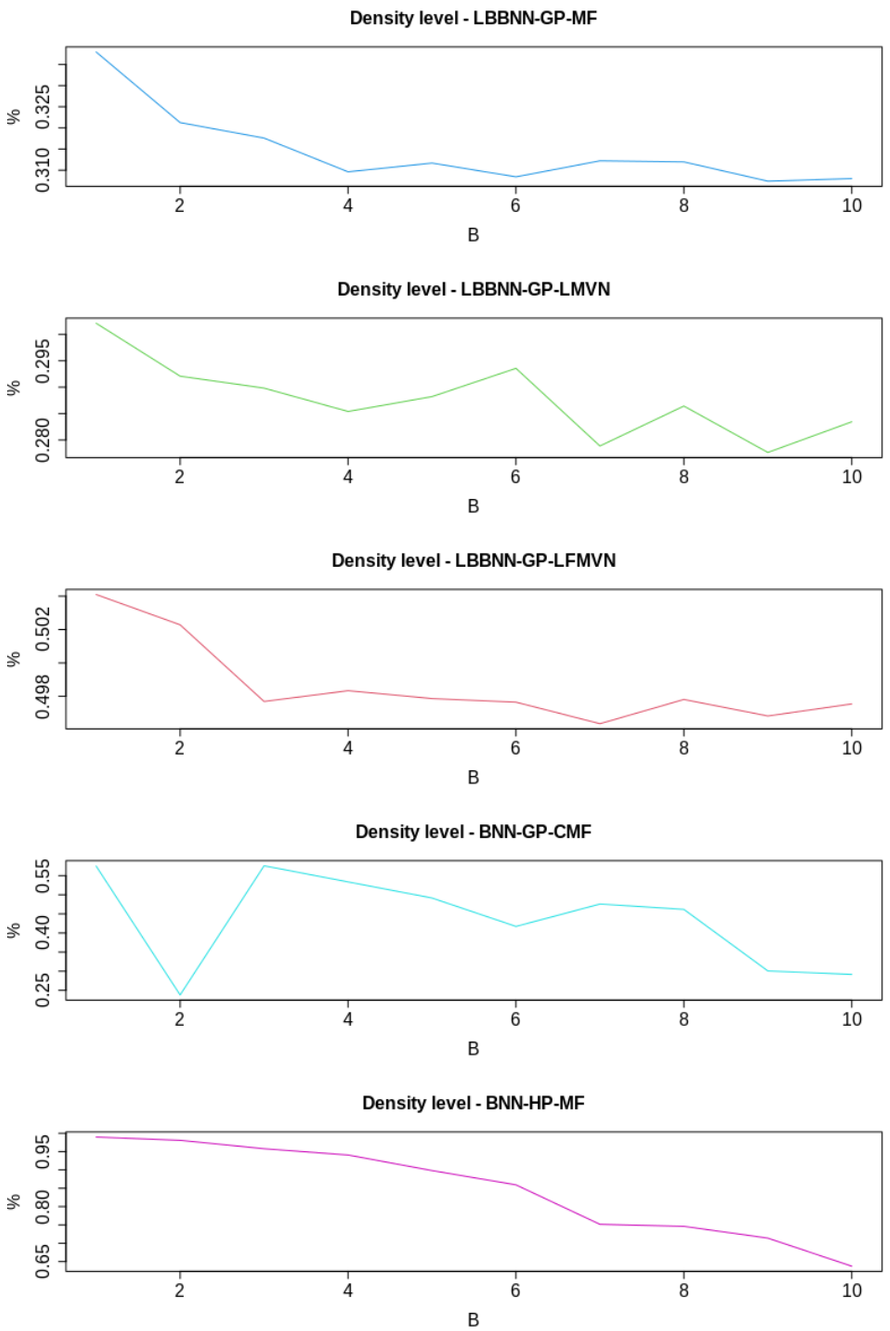

图13：不同 贝叶斯稀疏化方法在 PHONEMNE 数据集上不同层宽度的概率密度 a.

附录 F：PHONEMNE 数据中权重后验纳入之间相关结构的结果

本节报告以下内容的 1000 个样本之间的相关结构： $\gamma^{(3)}$ 它们来自各自的近似后验（选择第三层是因为其参数数量最少，但其他层的整体情况相同），这些模型为在 PHONEMNE 数据上训练的 LBBNN-GP-MF、LBBNN-GP-MVN 和 LBBNN-GP-LFMVN；使用的设置与正文第 4 节相同。结果见图 14-16，表明 LBBNN-GP-MF 具有近似对角相关结构，而正如预期，LBBNN-GP-MVN 和 LBBNN-GP-LFMVN 的结构丰富得多。

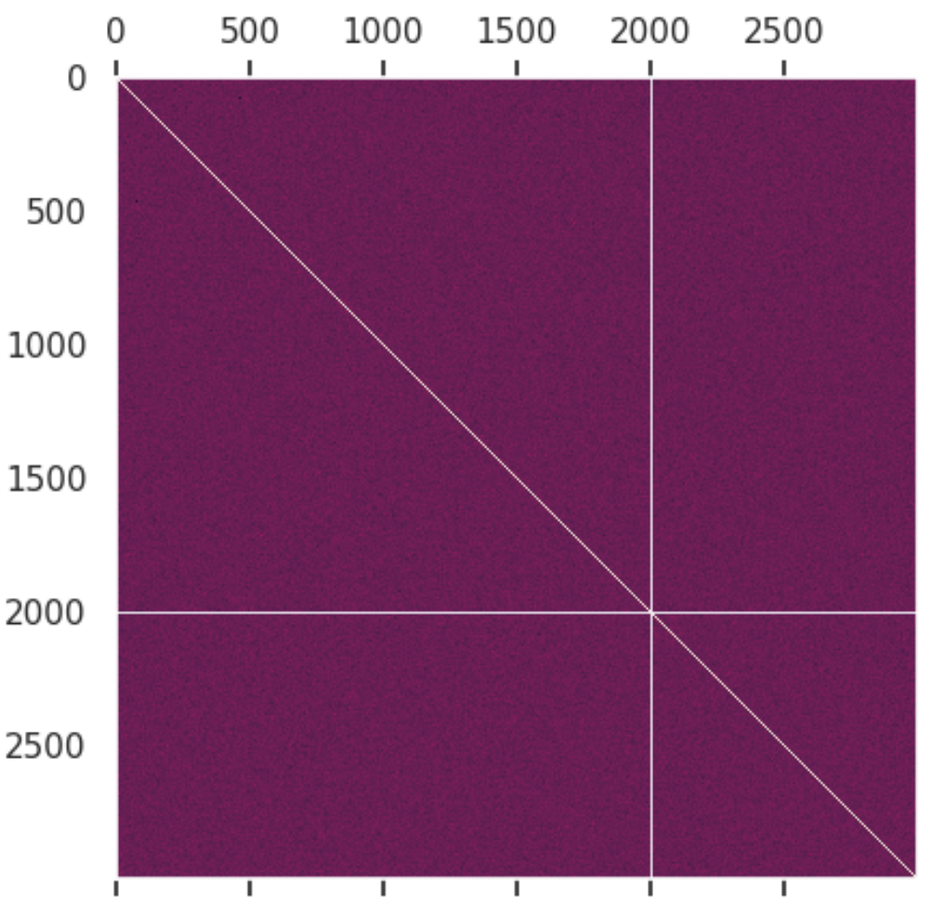

图14：PHONEMNE 数据上 LBBNN-GP-MF 第3层后验纳入指标之间的相关结构。

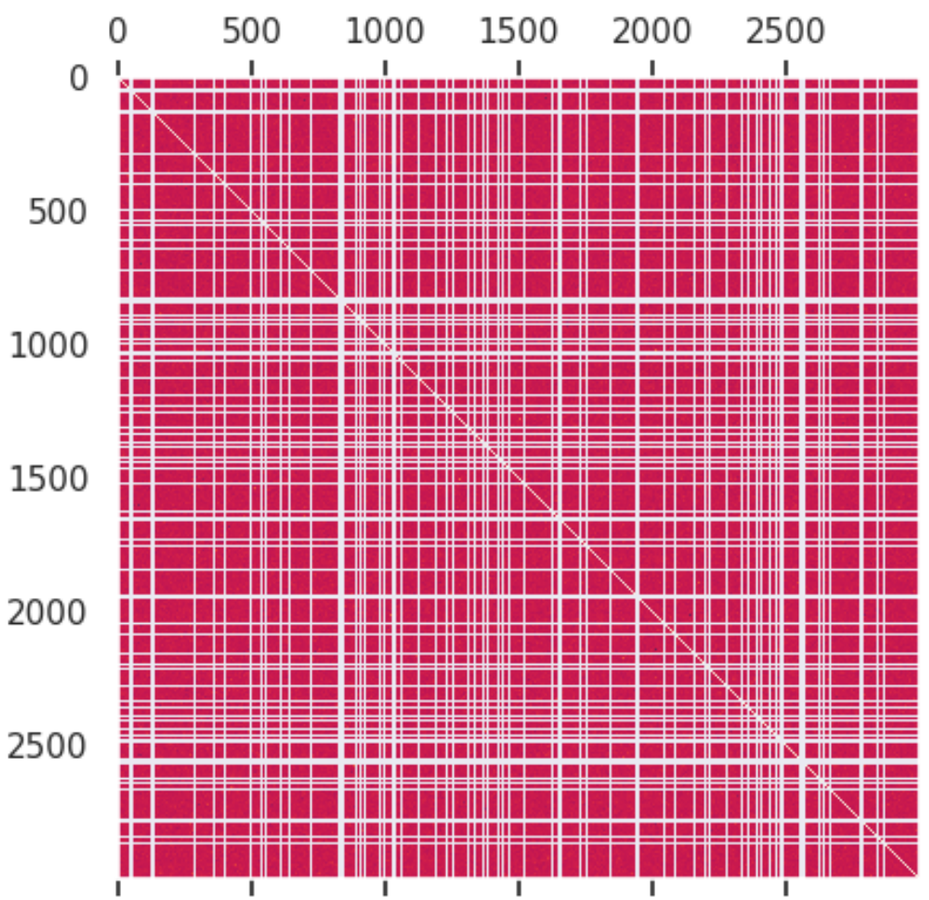

图1： PHONEMNE 数据中 LBBNN-GP-MVN 第3层后验纳入指标之间的相关结构。

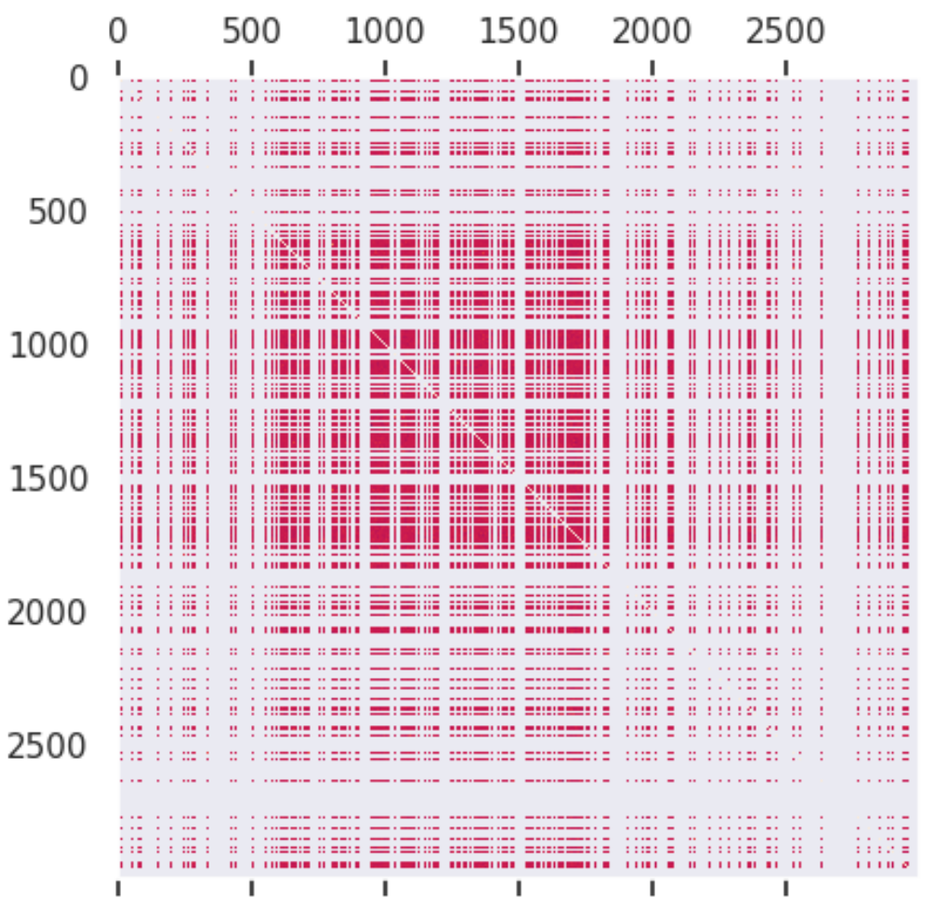

图16： LBBNN-GP-LFMVN 在 PHONEMNE 数据上第3层的后验纳入指标之间的相关结构。