高斯过程

# 常用符号表示

矩阵大写，向量为粗体。我们通常不区分概率和概率密度。下标星号，例如：，表示对测试集数量的引用。上标星号表示复共轭（complex conjugate）。

|  |  |
| --- | --- |
| Symbol | Meaning |
| \ | A\b为一个解决Ax=b的向量 |
|  | 作为定义的等式 |
|  | 等于一个附加常数 |
|  | K矩阵的行列式 |
|  | 向量y的欧几里得距离 |
|  | RKHS内积 |
|  | RKHS范数 |
|  | 向量y的转置 |
|  | 正比于 |
|  | 分布式根据 |
|  | 偏微分 |
|  | 海森矩阵的二阶导数 |
| C | 一个分类问题中的类数 |
| cholesky(A) | Cholesky分解：L为下三角矩阵 |
|  | 高斯过程后验协方差 |
| D | 输入空间X的维数 |
|  | 数据集： |
| diag(w) | 一个包含向量w的元素的对角矩阵 |
| diag(W) | 一个包含矩阵W的对角线元素的向量 |
|  | 克罗内克符号，如果p=q， |
|  | 期望当时，z(x)的期望 |
|  | 高斯过程(或矢量)隐（latent）函数值 |
|  | 高斯过程(后验)预测(随机变量) |
|  | 高斯过程后验均值 |
| h(x) | 固定基函数(或基函数集)或权函数 |
| H(X) | 所有训练点评估的一组基函数 |
|  | 单位矩阵(大小为n) |
|  | 第一类贝塞尔函数 |
|  | 在x和处的协方差函数 |
| K(X,X) | n×n协方差矩阵 |
|  | 训练和测试用例之间的协方差 |
|  | (无噪声)f值的协方差矩阵 |
|  | (有噪声的)y值的协方差矩阵； |
|  | 修正贝塞尔函数 |
|  | 损失函数，当a为真时，预测b的损失 |
|  | 特征长度尺度(用于输入维度d) |
|  | 逻辑函数， |
| m(x) | 高斯过程的均值函数 |
|  | 测量 |
|  | (变量x)的高斯分布，其中均值向量μ和协方差矩阵 |
|  | 训练和测试用例的数量 |
| N | 特征空间维度 |
|  | 神经网络中隐藏单元的个数 |
|  | 自然数，正整数 |
|  | 当时，保持有界，则有 |
|  | 零矩阵或者为微分算子 |
|  | 给定x条件随机变量y及其概率(密度) |
|  | 常规n-多边形 |
|  | 输入的特征图 |
|  | 累计单位高斯： |
|  |  |
|  | 地图预测：在评估π |
|  | 均值预测：的期望值 |
|  | 真实值 |
|  | 的风险或预期损失，或分类器c |
|  | 预测l的期望损失, |
|  | c的决策区域 |
| S(S) | 功率谱 |
|  | Sigmoid function |
|  | 无噪声方差 |
|  | 噪声方差 |
|  | 超参数向量(协方差函数参数) |
| tr(A) | 矩阵A的迹 |
|  | 周长为l的圆 |
|  | 当时，z(x)的方差 |
|  | 输入空间和随机过程的指标集 |
| X | 训练输入的D×n矩阵 |
|  | 测试输入矩阵 |
|  | 第i个训练输入 |
|  | 第i个训练输入xi的第d个坐标 |
|  | 整数 |

# 简介

通常我们用x表示输入，y表示输出，y可以是连续的也可以是离散的。我们有一个n观测的数据集D，。根据这些数据，我们希望对新的输入x进行预测。首先我们需要对基础函数特征进行假设，通常有两种方法：1. 限制函数类，例如只考虑输入的线性函数。2. 给定每个可能函数的先验概率，对于我们认为可能性更大的函数，给出更高的概率。

第一种扼杀了函数的丰富性，如果没有选对目标函数，预测效果会很差。但是如果增加函数类的丰富性，又会面临着过拟合的危险。

第二种方法函数存在无数种可能，是无法求解的。概率分布描述的是标量或者向量，这就是高斯过程发挥作用的地方。高斯过程是高斯概率分布的推广。虽然概率分布描述的是标量或向量(多变量分布)的随机变量，但一个随机过程（stochastic process）支配着函数的性质。

高斯过程框架的主要优点之一就是它将复杂一致的视图与计算可跟踪性结合起来。

# 回归问题

我们可以把高斯过程想象成在函数定义一个分布，然后直接在函数空间进行推理，及function-space view。

## Weight-space View

简单的线性回归模型，输出是输入的线性组合。它的主要缺点是只允许有限的灵活性，如果输入和输出之间的关系不能合理地用一个线性函数来近似，该模型将给出糟糕的预测。

我们有一个n个观察点的训练集合D，，其中x表示维度D的输入向量(协变量，covariates)，y表示标量输出或目标(因变量，dependent variable)；所有n种情况列向量的输入用D x n维矩阵X表示，目标集合在向量y中，因此我们可以写成D = (X, y)。在回归设定中，目标是实值。我们感兴趣的是推断输入和目标之间的关系，即给定输入的目标的条件分布。

### 标准线性模型

我们将回顾带高斯噪声的标准线性回归模型的贝叶斯分析：

（1）

其中，x是输入向量，w是权重向量(参数)的线性模型，是函数值，y是观测到的目标值。通常还会包含一个偏置权值（bias weight）或偏移量（offset），但是由于可以通过使用一个值总是1的附加元素来增加输入向量x来实现，所以我们没有在符号中显式地包含它。我们假设观测值y与函数值不同，函数值存在可加噪声，我们将进一步假设该噪声遵循一个独立的、同分布的高斯分布，其均值和方差均为零。

（2）

这个噪声假设与模型一起直接产生了似然，即给定参数的观测值的概率密度，在训练集中被分解的给出（独立假设）

 （3）似然

表示向量z的欧氏长度。在贝叶斯形式中，我们需要在参数之上指定一个先验，在观察之前表达我们对参数的信念（beliefs）。我们将协方差矩阵和一个零均值高斯先验加入到权重

（4）先验

贝叶斯线性模型的推论是基于权重的后验分布，由贝叶斯规则计算

（5）

其中归一化常数（normalizing constant）也称为边缘似然（marginal likelihood），与权重无关

（6）

式（6）中的后验结合了似然和先验，并捕获了我们所知道的关于参数的所有信息。只写出与权重有关的似然项和先验项，并进行平方，我们得到

（7）

其中，我们将后验分布的形式表示为带有均值和协方差矩阵的高斯分布。

（8）后验

其中。注意，对于这个模型(实际上对于任何高斯后验)，后验分布的均值也是其模态（mode），也称为的最大后验概率（maximum a posterior，MAP）估计。在非贝叶斯设置中，负对数先验（negative log prior）有时被认为是一个惩罚项，而MAP点被认为是惩罚权重的极大似然估计值。惩罚极大似然估计的一个典型例子是岭回归。

为了对测试用例进行预测，我们对所有可能的参数进行平均，加权后验概率。在的预测分布由对所有可能的线性模型的输出进行平均给定，关于高斯后验

（9）概率密度函数（pdf）回归

预测分布仍然是高斯分布，其平均值是由式（8）的权值的后验平均值乘以测试输入得出的，正如人们从对称性考虑中所期望的那样。预测方差是带有后验协方差矩阵的测试输入的二次形式，表明预测不确定性随测试输入的大小而增加，正如线性模型所期望的那样。

图1给出了贝叶斯回归的一个例子。

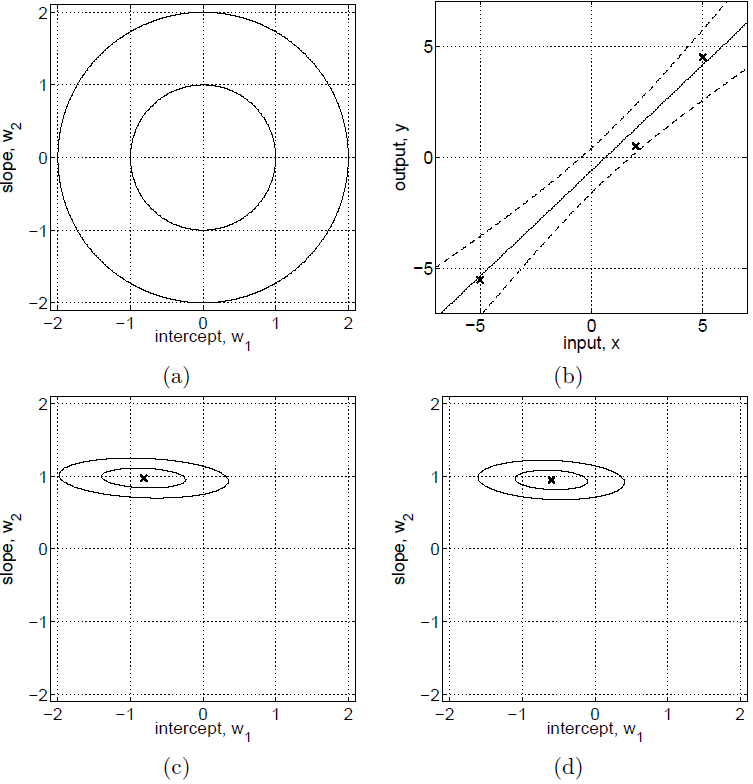


图1：贝叶斯模型，其中为截距，为斜率。（a）显示了先验分布的轮廓，参考式4。（b）显示了三个训练点，用\*表示。（c）似然线，参考式3，假设噪声水平为，可以看出斜率比截距确定的更好。（d）显示了后验，参考式7，将后验的最大值与似然进行比较，我们可以看到截距已经缩小到零，而斜率几乎没有变化。所有轮廓图给出1和2个标准差等概率轮廓。（b）中的两条虚线为预测分布的±标准差。

### 输入到特征空间的投影

该模型的后验与式8相同，这小节只是简化概率密度函数。

贝叶斯模型不能很好的处理高维模型，克服这个问题的一个方法是首先使用一组基函数将输入投影到某高维空间中，然后在高维空间中应用高维模型，而不是直接对输入本身。例如，标量输入x可以投影到x的幂空间（space of powers）中：，实现多项式回归。只要投影是固定的函数(即与参数w无关)，模型在参数中仍然是线性的。

特别的，我们引入函数，其将D维输入向量x映射到N维特征空间。进一步让矩阵为列的聚合。现在模型为

（10）

参数向量的长度现在是N。这个模型的分析类似于标准线性模型，不同的是被X替代。这样预测分布就变成了

（11）

其中。

为了利用这个方程进行预测，我们需要对一个N×N的矩阵求逆，如果特征空间的维数N很大，这可能不太方便。因此重写方程。

（12）概率密度回归

其中。为了表示均值，首先注意使用A和K的定义，我们有。变换得。均值参考式11和式12，对于方差我们矩阵求逆引理，式13。设，其中。在式12中，我们需要对n x n矩阵求逆。

（13）

在式12中，特征空间总是以进入。因此这些矩阵的元素形式总是不变为，其中和在训练集或者在测试集中。让我们定义。我们将k称为协方差函数或者是核函数（kernel）。注意：是一个内积（关于的）。因为是正定的，我们可以定义为。例如，奇异分解（singular value decomposition，SVD），其中D是对角线矩阵，对应。定义，我们获得一个简单的点积表示。

## Function-space View

定义：高斯过程是随机变量的集合，任何有限个数都有一个联合高斯分布。

一个高斯过程完全由它的均值函数和协方差函数来确定。定义均值函数和协方差函数为

（14）

高斯过程可以写为

（15）

通常，为了标记的简单性，我们将取均值函数为零。

在我们的例子中，随机变量表示函数f(x)在位置x处的值。通常，高斯过程是在时间上定义的，即随机变量的索引集是时间。

高斯过程首先假设函数服从一个先验分布，然后根据观测数据来约束这个分布，即基于每个函数与观测样本的相关程度，找出最能反映观测样本特征的函数，进行更可靠的预测。

高斯过程具有一些性质。例如如果，那么，其中是的相关子矩阵，如式16所示。



一个简单的高斯过程例子可以由贝叶斯线性回归模型得到，其中先验为，可得均值和方差

（17）

因此和是由给定的具有零均值和协方差联合高斯分布，函数值是对应任意数量的输入点n的联合高斯分布。如果，高斯是奇异的。

协方差函数指定随机变量对之间的协方差

（18）

注意，输出之间的协方差被写成输入的函数。平方指数协方差函数对应于具有无穷个基函数的贝叶斯线性回归模型。事实上，对于每个正定的协方差函数k，在基函数方面存在（可能无限的）扩展。我们可以通过无限个高斯基函数的线性组合得到平方指数（squared exponential，SE）协方差，如式19和20所示。

（19）

（20）

协方差函数的指定意味着函数上的分布。为了了解这一点，我们可以从函数在任意点的分布中抽取样本，其中，我们选择了一些输入点，用式18的elementwise写出相应的协方差矩阵。然后我们用这个协方差矩阵生成一个随机高斯向量

（21）

如图2(a)所示显示了三个这样的样本。输入值是等距的，但是实际情况下这又是不可能的。平方指数协方差函数是无限可微的，从而导致过程是无限均方可微的。我们也看到函数似乎有一个特征的长度尺度，可以认为是在函数值发生显著变化之前，必须在输入空间中移动的大致距离。对于式18，特征长度尺度约为一个单位。将式18中的替换为，这样就可以改变过程的特征长度尺度。此外，式18中exp前的一个正前因子（positive pre-factor）可以控制随机函数的总方差。我们将在下节讨论这些因素如何影响预测。

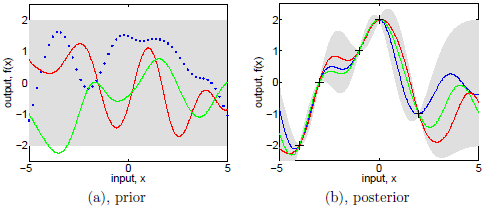


图2：(a)显示从GP先验随机绘制的三个函数，圆点表示实际生成的y值;另外两个函数通过连接大量的计算点被画成直线；(b)显示了从后验中提取的三个随机函数，即以所示的五个无噪声观测值为条件的先验函数。在这两个图中，阴影区域代表了每个输入值(对应于95%置信区间)的点态平均值加减两倍的标准差，分别表示先验和后验。

### 通过无噪声观测进行预测

首先，我们将考虑无噪声观测的简单特殊情况，我们已知。根据先验训练集输出联合概率分布和测试集输出的。

（22）

如果有n个训练点，个测试点，则定义了一个的在所有成对的训练和测试点上评估的协方差矩阵。为了得到函数的后验分布，我们需要限制这个联合先验分布只包含那些与观测数据点一致的函数。从图2中可以形象地看出，我们可以考虑从先验生成函数，并排除与观察结果不一致的函数。这种策略在计算上不是很有效。幸运的是，从概率的角度来说，这个操作是非常简单的，它对应于在观测值上设置联合高斯先验分布的条件

（23）后验，回归

函数值,可以通过评估均值和方差矩阵从联合后验分布中采样。图2(b)给出了5个数据点的计算结果。请注意，将这些计算扩展到多维输入是非常简单的，只需根据式18更改协方差函数的评估，尽管结果函数可能更难以图形化地显示。

### 使用有噪声的观测进行预测

设，增加独立同分布高斯噪声，方差为，噪声观测的先验变为

（24）

其中如果p=q，则为1，否则为0。根据噪声的独立性假设，在无噪声的情况下，参考式18，加入一个对角矩阵。在式22中引入噪声项，可以将观测到的目标值与在先验下的测试点处的函数值的联合分布写成

（25）

对式23推导对应的条件分布，得到高斯过程回归的关键预测方程

（26）

其中

（27）

（28）

注意，如果定义我们现在与weight space中式12完全一致，其中C,D代表X或。对于任何一组基函数，我们可以计算相应的协方差函数为。相反对于每一个（正定）协方差函数k，在基函数方面存在一个（可能无限的）展开式。我们可以将式27和28写成简单表达形式

（29）均值后验

（30）方差后验

式29和20给出了预测分布，首先注意，均值预测式29是观测值y的线性组合，被称为线性预测。另一种理解这个方程的方法是把它看成n个核函数的线性组合，每个都以一个训练点为中心

（31）

其中。的预测均值可以写作式31，尽管GP可以用一组(可能是无限的)基函数来表示，这仍然是表示定理（representer theorem）的一种表现形式。我们可以直观地理解这个结果，因为尽管GP在所有y变量上定义了一个联合高斯分布，在索引集合中的每个点，在做预测，我们只关注由n个训练点和测试点的n+1维分布。由于高斯分布仅通过取联合协方差矩阵的相关块而被边缘化，很明显，将这个(n+1)维分布限定在观测值上就得到了我们想要的结果。图3给出了GP的图形模型表示。

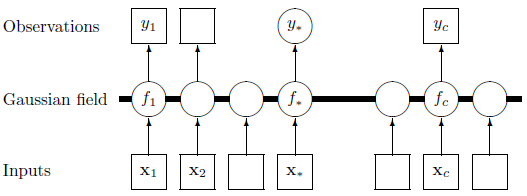


图3：正方形代表观察到的变量，圆形代表未知数。粗横杠表示一组完全连接的节点。注意一个观测值在给定相应隐变量的情况下，独立于其他所有节点。由于GPs的边际属性，增加了进一步的输入，x，潜在变量，f，和未观测到的目标，，不改变其他变量的分布

仍需注意的是，式28中的方差不依赖于观测到的目标，只依赖于输入，这是高斯分布的一个性质。两项之间的方差是不一样的：第一项是简单的先验方差，再从中减去一项(正的)，表示观察到的关于函数的信息。通过增加到，我们可以很简单的计算测试目标的预测分布。

GP模型的预测分布给出的不仅仅是简化公式30的pointwise errorbars。虽然没有明确说明，但当表示多个测试输入式28保持不变。在这种情况下，计算测试目标的协方差(其对角线元素是pointwise方差)。事实上，式27是高斯后验过程的均值函数和28是协方差函数，如图4所示。

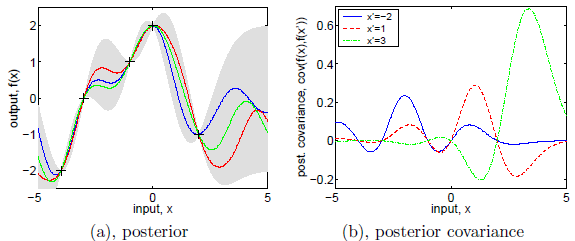


图4：(a)与图2(b)相同，显示了从后部抽取的三个随机函数。(b)显示了三个不同值的和之间的后验协方差。注意，闭点的协方差很高，在训练点(没有方差，因为这是一个无噪声的过程)降为零，然后变为负值。发生这种情况是因为，如果平滑函数恰好小于数据点一侧的均值，那么它往往会超过另一侧的均值，从而导致数据点处协方差符号的反转。注意，作为对比，先验协方差是简单的高斯形状，不为负。

这对边缘似然（marginal likelihood）是非常有用的，边缘似然是先验乘以似然的积分

（32）

边缘似然是指函数值f的边际化，在高斯过程模型下，先验是高斯的，，或者为

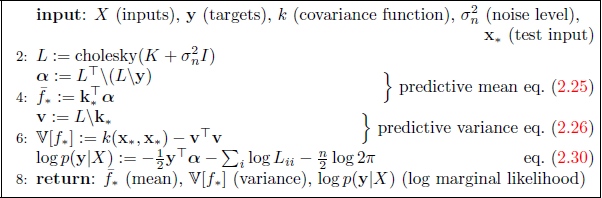
（33）

似然是一个映像高斯，所以我们来执行产生对数边缘似然的积分

（34）

这个结果也可以通过观察直接得到。

算法1给出了高斯过程回归（GPR）的一个实际实现。该算法使用Cholesky分解，而不是直接对矩阵求逆，因为它更快，在数值上更稳定。该算法返回无噪声测试数据的预测均值和方差，计算有噪声测试数据的预测分布。简单的将噪声方差加到预测方差。



34

30

29

## 改变超参数（Hyperparameters）

通常我们使用的协方差函数会有一些自由参数。例如，一维的平方指数协方差函数有以下形式

（35）

协方差记为，表示有噪声的目标y，而不是底层函数f，观察其长度尺度，信号方差和噪声方差能够被改变。通常我们称自由参数为超参数。（我们把协方差函数的参数称为超参数，以强调它们是非参数模型的参数）

如图5所示，这是从带有SE内核以及超参数的GP生成的。当输入值远离任何训练点时，错误条会变得更大。如果我们将长度尺度设置得更短，使并保持其他参数不变，那么从这个过程中生成的图将与图5(a)类似，只不过x轴应该重新调整0.3倍，同样地，如果保持图5(a)中相同的x轴，则示例函数看起来会更加波动。

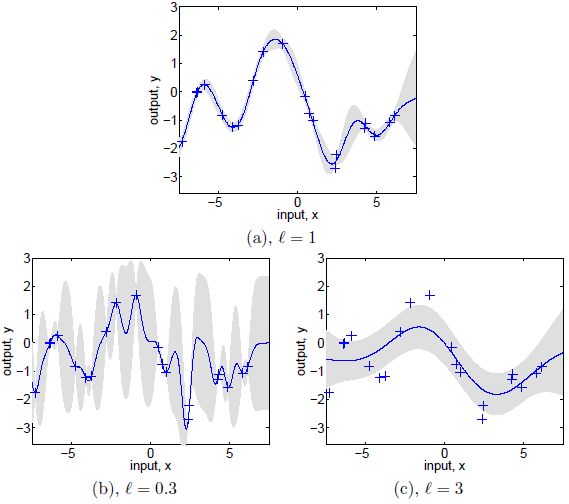


图5：数据来自具有超参数的GP，利用这些超参数进行高斯过程预测，我们得到了底层函数f的95%置信区间(灰色)所示。(b)和(c)再次显示了95%置信区间，但这次是超参数值(0.3、1.08、0.00005)和(3.0、1.16、0.89)。

## 回归决策理论

预测分布为高斯分布，其均值和方差由式29和式30给出。然而，在实际应用中，我们经常被迫决定如何行动，也就是说，我们需要一个在某种意义上是最优的点状（point-like）预测。为此我们需要一个损失函数，损失函数可以等于猜测与事实之间的绝对偏差。

注意，我们在计算预测分布时没有参考损失函数。在非贝叶斯范式中，模型通常是通过最小化经验风险(或损失)来训练的。相比之下，在贝叶斯设置中，在似然函数(用于训练，加上先验)和损失函数之间有一个清晰的分离。似然函数描述了如何假设噪声测量值来使其偏离底层噪声函数。另一方面，损失函数在给定真实状态的情况下，捕捉做出特定选择的结果。可能性和损失函数可以没有任何共同点。

我们的目标是预测会导致最小的损失，但是当我们不知道是真的时候，我们怎么能做到呢？我们最小化预期损失或风险通过平均我们模型的意见。

（36）

因此我们在最小化预期损失的意义上最好的猜测为

（37）

通常的值是的中值，使损失函数的风险最小化，对于损失函数，它是分布的均值。当预测分布为高斯分布时，均值和中值重合，对于任何对称损失函数和对称预测分布，我们总是得到作为预测分布的均值。然而，在许多实际问题中，损失函数可能是不对称的。

## 用例说明

使用高斯过程回归来学习七自由度SARCOS拟人化机器人手臂逆动力学。这个任务从一个21维的输入空间（7个关节位置，7个关节速度，7个关节加速度）到相应的7个关节扭矩。

为什么要学系这种映射，确实存在基于物理的刚体动力学模型允许我们从位置、速度和加速度变量中获得力矩。然而，真正的机器人手臂是液压驱动的，而且相当轻量和顺从，所以刚体动力学模型的假设被违反。值得注意的是，刚体动力学模型是非线性的，包括三角函数和输入变量的平方。

逆向动力学模型可以以以下方式使用：规划模块决定将机器人从开始状态带到目标状态的轨迹，并指定每次所需的位置、速度和加速度。反动力学模型用于计算实现该轨迹所需的力矩，并使用反馈控制器对误差进行修正。

数据集由48933输入-输出对组成，其中44484个被用作训练集，其余4449个作为测试集，输入被线性调整为零均值和单位方差，输出居中使均值为零。

给定一种预测方法，我们可以从几个方面来评估预测的质量。也许最简单的是平方误差损失，其中我们计算每个测试点均值的预测和目标之间的平方残差（residual）。这可以通过均方误差(MSE)平均测试集总结。然而,这个数量对目标值的总体规模很敏感，因此通过测试用例目标的方差进行规范化，以获得标准化的均方误差(SMSE)，这是有意义的。这导致猜测训练目标均值的简单方法的SMSE大约为1。

此外，如果我们在每个测试输入处生成一个预测分布，我们就可以评估模型下目标的负对数概率。当GPR产生一个高斯预测密度时，我们得到

（38）

其中GPR的预测方差为，其中由式30给定。我们在预测噪声目标时必须包括噪声方差。这种损失可以通过减去在平凡模型（trivial model）下得到的损失来标准化，平凡模型利用训练数据的均值和方差采用高斯分布进行预测。我们称其为标准化日志损失(SLL)。均值SLL记作MSLL。对于简单的方法，MSLL近似为零，对于更好的方法，MSLL为负。