arXiv:2304.01120v1 [cond-mat.mtrl-sci] 3 Apr 2023

利用定量表征和高维分布距离进行合成参数效应检测

Alex Hagen、Shane Jackson

摘要

检测合成过程参数对材料微观结构的影响，是材料科学中的一项重要但难以实现的目标。我们提出了一种基于 Copula 理论、高维分布距离和置换统计的效应检测方法，用于分析一项以草酸钚(III)合成氧化钚的设计实验。我们检测到冲击顺序和草酸进料对所得氧化钚微观结构的影响，这些结果与文献高度吻合。我们还检测到酸浓度、冲击顺序和沉淀温度两两之间存在额外的双变量效应。

检测材料因组分、加工步骤、加工环境及其他材料历史差异而产生的变化，是材料科学所需的一项核心能力。这一过程统称为“效应检测”，是新材料发现 [1], 材料性能优化 [2],材料来源特征发现 [3], 以及其他广义材料性能研究的第一步。在本研究中，我们将高维分布统计应用于材料效应检测这一挑战，其中响应变量为微观结构。为此，我们将卷积神经网络应用于合成颗粒的显微图像，生成对成像微观结构的定量描述，随后开展高维且统计严谨的效应检测。将效应检测从宏观性能拓展至微观结构，将对材料科学产生广泛影响，涵盖设计实验结果分析、数据驱动的来源分析以及自动化材料发现。

材料科学中关于效应检测的文献十分丰富，尽管主要集中于方差分析（ANOVA）方法的各种变体 [4]。这些方法需要满足两个限制条件：响应变量必须是一维的，且底层分布能够用高斯分布很好地近似。例如，材料的微观结构（如生成锕系元素 [5], 的过程分析中的微观结构）具有学术和实践价值。在这类情形中，效应检测较为复杂，因为微观结构是颗粒的一种定性特征。通常，人们会借助某种显微术来评估颗粒的微观结构，例如二次电子扫描电子显微镜（SEM）。一些研究者尝试对这类微观结构进行分类，以便将其纳入定量框架 [6]。另一些研究者借鉴自然图像计算机视觉这一广阔领域的方法，使用卷积神经网络对SEM显微图像进行定量描述 [7, 8, 9]。我们采用了后一种技术，因为它具有良好的可重复性且速度快。

要将定量描述方案扩展到效应检测，我们需要一种统计检验。在这类显微图像的定量描述方案中，输出描述具有高维特性。因此，我们需要对两个高维变量进行独立性检验。在我们的综述和实证测试中，用于高维依赖关系的显式检验效果不佳——例如，距离相关和协方差 [10]难以检测非线性或非单调关系 [11, 12], ，而且在我们的实验中，它们要么返回异常结果[1], ，要么在 Copula 表述中 [11] 产生许多假阳性[2]。因此，我们转而采用新方法来检测高维效应。

[1]例如，相关性可证明小于两个随机变量之间的相关性，实际上是真正的负相关[2]我们将假阳性定义为检测到两个实际上随机的向量或矩阵之间存在效应

表 1：Pu(III)　草酸盐沉淀合成参数及其对应的标签编码。经 [16]许可转载。
许可。

	参数	编码：0	1	2
A	草酸进料	0.9 M 溶液	固体	-
B	添加时间（分钟）	0	20	40
C	消解时间（分钟）	40	20	0
D	HNO_3 浓度（M）	1.0	2.0	3.0
E	冲击顺序	反向	正向	-
F	沉淀温度（°C）	30	50	-
G	钚浓度（g/L）	10	30	50

在一维情况下，统计距离很容易定义；然而，在更高维度下，累积分布函数定义上的模糊性使距离存在多种可能的定义[3]。此外，由于计算成本较高，许多方法依赖于对底层分布的假设；如果对数据集的先验知识$a\ priori$ 有限，这些假设可能会带来问题。例如，我们的方法有一种类似方法，即 Frechet Inception Distance（FID） [13], ，常用于自然图像研究文献中比较合成图像与真实图像。该距离的基础假设是所有分布均为高斯分布，因此可以仅通过比较矩来进行比较；对于许多与材料科学相关的数据集，这一假设很容易被证伪。因此，我们的研究重点是非参数高维距离，例如 d维柯尔莫哥洛夫–斯米尔诺夫距离（ddks） [14]和最大均值差异（mmd）[15] 。

我们研究了一组由草酸钚(III)煅烧生成 PuO_2，的显微图像，这些样品是在一项包含 76 次运行的设计实验 [16] 中制备的。该设计实验改变了 7 个参数的取值：(A) 草酸进料、(B) 添加时间、(C) 消解时间、(D) HNO_3 浓度、(E) 冲击顺序、(F) 沉淀温度和 (G) 钚浓度。它们的可能取值列于表 1。在 76 次运行中，有 50 次符合质量控制要求，并纳入本研究。我们首先从整体显微图像中分离出每个颗粒，然后使用 [7] 中的矢量量化变分自编码器（VQVAE）对其进行编码，以此对数据进行预处理。

我们开发了一种基于分布距离和扰动的方法，通过概率积分变换来检测参数对高维响应的影响。给定 N 个具有输入参数 $\vec{x} = [x_0, x_1, \ldots, x_N]^T$ 和响应 $\vec{y} = [y_0, y_1, \ldots, y_N]^T$ 的不同颗粒，其中 $x \in \mathbb{R}^{d_x}$ 和 $y \in \mathbb{R}^{d_y}$ 的元素。

在效应检测中，我们的目标是识别并量化输入参数 $\mathcal{X}$ 与响应变量 $\mathcal{Y}$ 的概率分布。为此，我们考虑联合参数 $\vec{z} = (\vec{x},\ \vec{y})$ 的概率分布，并记为 $\mathcal{Z}$。通过置换 $\vec{x}$ 的各行，同时保持 $\vec{y}$ 不变，可生成一个$\tilde{\vec{z}}$第二联合参数 Z，它具有条件独立性。通过双样本检验来比较 $Z\tilde{\vec{z}}$与 Z，例如 d 维 Kolmogorov-Smirnov（ddKS）检验 [14], 或最大均值差异（ MMD ）检验 [17]。

实际上，我们可以通过将来自 $\vec{x}$ 和 $\vec{y}$ 的样本分别拼接到数据矩阵 $\mathbf{X}$ 和 $\mathbf{Y}$ 中，构造一个维度为 $N \times (d_x + d_y)$ 的数据矩阵，并将拼接结果记为 $\mathbf{Z}$。我们通过从数据矩阵中随机选取行来生成联合分布的样本。我们通过从 $\mathbf{Z}$ 中抽取行并仅随机置换 $\mathbf{X}$ 的各行，来生成联合独立分布的样本；等价地，也可以置换数据矩阵中与 $\mathbf{X}$ 相关列的各项。我们将置换后的数据$\tilde{}$矩阵记为 $\mathbf{Z} \equiv (\pi(\vec{x})\ \ \vec{y})$，其中 $\pi(\cdot)$ 表示随机逐行置换。

为了确定$\mathbf{Z}$与$\tilde{\mathbf{Z}}$是否来自同一个分布时，我们使用前述距离来衡量它们的分布距离。一般而言，这些距离仅受限于为正

[3]在二维情况下，CDF 可以定义为 $F = P(X < x, Y < y)$ 或 $F(X < x, Y > y)$，也可以采用其任意置换。

实数，且给定距离的统计显著性可能未知。我们使用置换检验 [18] 计算给定距离的精确统计显著性。这种显著性实际上成为效应检测的衡量指标：如果参数与响应相互独立，则显著性以 $1-\alpha$ 的置信度低于阈值 $1-\alpha$；否则，响应依赖于参数。

上述效应检测方法可扩展至多个合成参数。朴素的扩展方式~会将 $\vec{x}_1$、 $\vec{x}_2$ 和 $\vec{y}$ 的联合分布与随机打乱后合并的数据集 **Z**进行比较。这种扩展测量的是 $\vec{x}_1$ **或** $\vec{x}_2$（此处“或”包含两者同时成立的情况）的效应。相较之下，材料科学家更关注 $\vec{x}_1$ **和** $\vec{x}_2$ 的效应，即无法单独由任一参数解释的效应。

我们首先计算 $\vec{x}_1$ 和 $\vec{x}_2$ 的经验累积分布函数，并使用逆概率变换将其转换为均匀分布变量 U_1和 U_2。然后，$\mathbf{Z} \equiv (U_1,\ U_2,\ \vec{y})$与置换后的 $\mathbf{Z} \equiv (\pi(U_1,\ U_2)\ \vec{y})$之间的分布~距离，可用于衡量 U_1 和 U_2的综合效应。为确定该距离的抽样分布，我们还可以在分别置换 U_1 和 U_2 时进行~ ~置换检验。我们将这些距离记为 d、 d_1和 d_2。

如果 $\vec{x}_1$ 和 $\vec{x}_2$ 高度相关，上述距离容易产生假阳性。为减轻这一影响，我们首先计算 $\vec{x}_1$与 $\vec{x}_2$之间分布距离差异显著性的补数。该值与 $\vec{x}_1$ 和 $\vec{x}_2$之间的相关性成反比，我们将其称为 r。我们可以通过两种不同方式对 r 进行置换检验：置换 $\vec{x}_1$ 和置换 $\vec{x}_2$，分别记为 $\tilde{r}_1$和 $\tilde{r}_2$。

~ ~将 r 与 d 的乘积同 $\tilde{r}_1 \cdot d_1$ 与 $\tilde{r}_2 \cdot d_2$ 的乘积进行比较，可直接将检验统计量 $r \cdot d$ 与其抽样分布进行比较。可以通过比较 $r \cdot d$ 与$\tilde{r} \cdot \tilde{d} =$最大值 $\left(\tilde{r}_1 \cdot \tilde{d}_1, \tilde{r}_2 \cdot \tilde{d}_2\right)$来构造检验统计量。随后，通过随机抽样对中满足 $r \cdot d > \tilde{r} \cdot d$的比例~来表示多参数效应的显著性。

我们使用两种不同的分布距离——最大均值差异和 ddKS，实施了上述方法。每次计算使用的样本量从 10 张显微图像到 5000 张显微图像不等。每次计算的置换操作进行了 100 次，并将所有计算重复 25 次，以考虑随机机会。结果如图 1 所示。

此前， [16]已对各参数的影响进行了研究。研究使用分类器将每个颗粒的显微图像归入各项属性的不同取值类别。各参数所达到的最高准确率依次为 94.1%、91.4%、91.4%、84.9%、98.0%、88.8% 和 80.6%。利用 [19], 的框架，我们可以将 [16] 中的分类器扩展为双样本检验，并确定达到显著性所需的样本数 α。这样便可将 [16] 与 ddKS 和最大均值差异直接比较。我们在附录 A 中推导所需的样本数，并将该计算的结果绘制于图 1 中。

图 1 中所有指标所显示的趋势是一致的。在二元参数中，使用最少样本量即可检测到参数 E（冲击顺序）对微观结构的影响，其次是参数 A（草酸进料），最后是参数 F（沉淀温度）。在三元参数中，排名逐渐升高的顺序为 B/C（加入/消解时间——互补变量）、D（HNO_3浓度），最后是 G（钚浓度）。

采用多参数效应方法研究形貌上是否存在双变量效应。样本量选取为 5000，以确保具备足够的统计功效来检测二阶效应。由于在该样本量下 MMD 所需内存过大，且 CB2S 不适用于多参数情况，因此我们仅使用 ddKS 进行研究。我们对所有变量对之间的比较进行了 500 次试验。结果如图 2 所示。酸浓度与沉淀温度之间、酸浓度与冲击顺序之间，以及冲击顺序与沉淀温度之间均存在显著的多参数效应。尽管这些效应在物理上是合理的，但在开展更多实验之前，我们没有其他验证机制。

这些结果表明，使用我们的方法进行的分析与 [16],的分析完全一致，而且由此展开的材料科学相关讨论比本文中的讨论更为深入。采用不同的统计方法（尽管使用的是同一个图像分析网络）对这些结果进行验证，这一事实共同增强了我们的方法及其结果的可信度。

我们的工作探索了广义意义上的多种效应检测方法，不假设

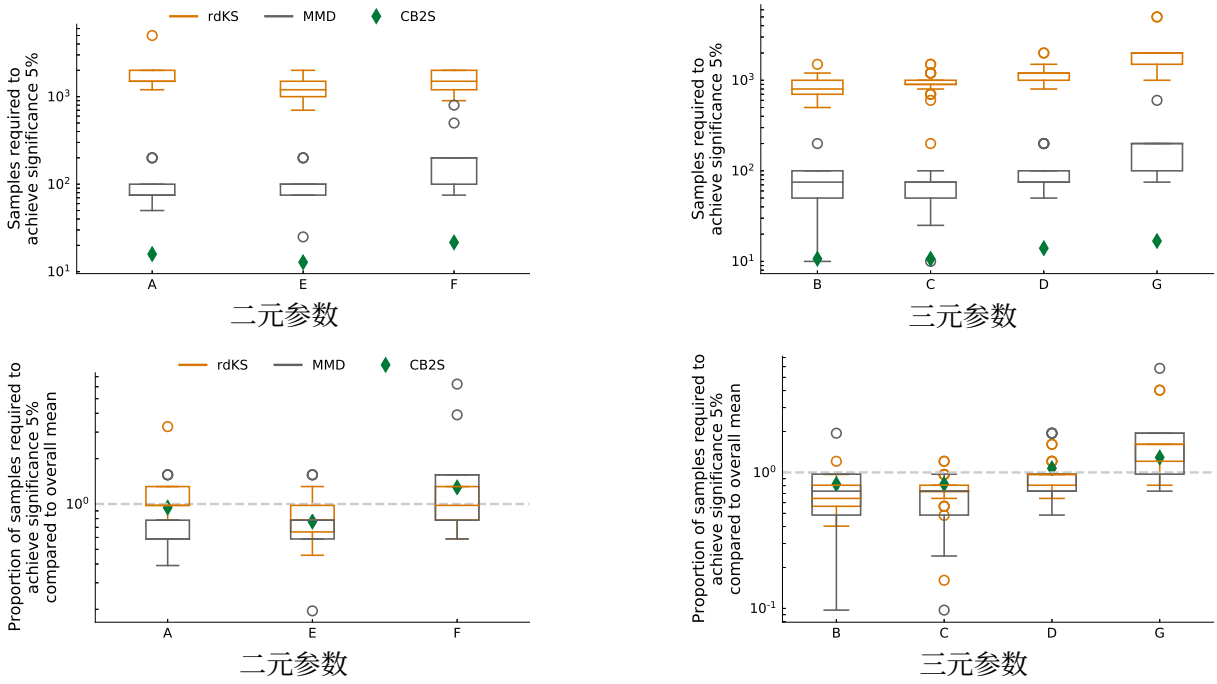

图 1：（第一行）：使用我们的方法以及分布距离 ddKS、最大均值差异和来自 [16]的基于分类器的双样本检验（CB2S）时，达到 95%显著性所需的编码显微图像数量。（第二行）：将上述计算结果除以各方法达到 95%显著性所需的平均样本数，表明 ddKS、最大均值差异和 CB2S 检测到的趋势相同，但统计功效有所不同。

参数或响应是一维的。我们证明，两个参数之间的距离相关独立性度量存在一些不理想的性质（例如，在其最初的应用中会出现负距离，或在其 Copula 表述中产生假阳性）。我们提出了一种方法来避免这些问题，并基于两种不同的底层距离度量取得了结果。

我们将效应检测方法扩展到多个参数，并研究了数据集中的双变量效应。我们将基于高维距离度量的效应量检测方法推广为一种仅检测由一对参数产生的超额效应量的方法。我们发现，酸浓度、冲击顺序和沉淀温度这几组参数对均存在显著的超额双变量效应。目前正在进一步研究这些效应背后的物理学原理和材料科学问题。

虽然不适用于多参数情形，但我们注意到，在单变量情形下，有几个方面使基于分类器的双样本检验更具优势。如图1所示，CB2S比基于距离的方法更强大，仅需少几个数量级的样本即可检测到相同的效应。此外，CB2S不依赖于线性效应，甚至也不依赖于单调效应。文中介绍的ddKS方法同样适用于非线性、非单调效应；但基于MMD的方法未能检测出与每组独特合成参数相关联的任意编号“运行”编号，这表明该方法并不适用于此类效应。最后，与基于距离的方法相比，CB2S能够满足更多效应检测的期望。在一般情况下，效应量描述了一种复杂关系，其中包括微观结构变化的程度，以及表现出这种变化的颗粒比例。基于距离的方法仅报告两个分布之间的距离，因此无法将两种可能的效应表现形式区分开来。CB2S允许查询单个颗粒，因此能够区分这两种表现形式。

尽管对本研究所检测效应开展实验研究构成了大部分可能的后续拓展方向，但改进本文提出的方法也会很有价值。ddKS 和最大均值差异距离度量的计算量都很大，因此采用其他距离度量或可降低

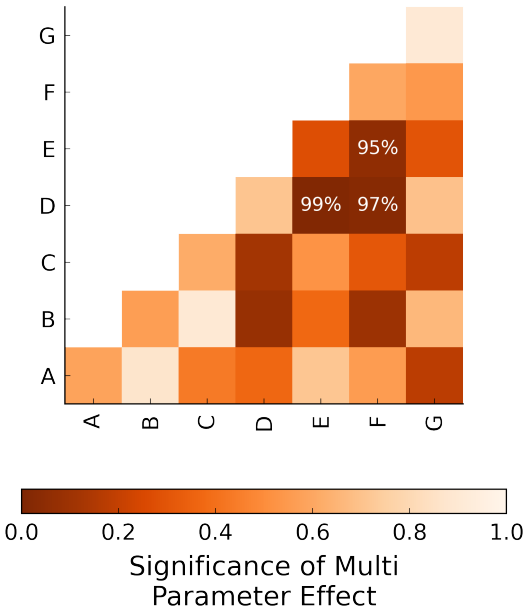

图 2：数据集中所有唯一变量对之间双变量效应的差异显著性。D（HNO_3 浓度）与 E（冲击顺序）、D 与 F（沉淀温度）以及 E 与 F 均达到 5% 显著性水平。

计算成本。此外，本文提出的多参数效应方法具有超多项式时间复杂度，随着问题规模扩大，其计算成本和所需样本量都会变得难以承受；因此，若能找到替代方法来缓解这些限制，将很有价值。

参考文献

[1] T. M. Belgasam、H. M. Zbib，影响双相钢能量吸收的关键因素：多尺度材料模型方法与微观结构优化，Metallurgical and Materials Transactions A 49（2018年）2419–2440。doi:10.1007/s11661-018-4563-x。[2] T. M. Belgasam、H. M. Zbib，使用代表性体积单元和响应面法优化双相钢微观结构：参数研究，Metallurgical and Materials Transactions A 48（2017年）6153–6177。doi:10.1007/s11661-017-4351-z。[3] T. Burr、I. Schwerdt、K. Sentz、L. McDonald、M. Wilkerson，使用颗粒形貌进行爆前核取证的算法概述，Algorithms 14（2021年）340。 doi:10.3390/a14120340。[4] P. Moran、C. Smith，孟德尔遗传假说下亲属间的相关性，Transactions of the Royal Society of Edinburgh 52（1918年）899–438。[5] E. Keegan、M. J. Kristo、M. Colella、M. Robel、R. Williams、R. Lindvall、G. Eppich、S. Roberts、L. Borg、A. Gaffney、J. Plaue、H. Wong、J. Davis、E. Loi、M. Reinhard、I. Hutcheon，在澳大利亚刑事调查中查获的未知铀矿石精矿样品的核取证分析，Forensic Science International 240（2014年）111–121。doi:10.1016/j.forsciint.2014.04.004。[6] A. L. Tamasi、L. J. Cash、C. Eley、R. B. Porter、D. L. Pugmire、A. R. Ross、C. E. Ruggiero、L. Tandon、G. L. Wagner、J. R. Walensky、A. D. Wall、M. P. Wilkerson，核取证中材料图像一致性描述词典，Journal of Radioanalytical and Nuclear Chemistry 307（2016年）1611–1619。 doi:10.1007/s10967-015-4455-0。[7] M. Girard、A. Hagen、I. Schwerdt、M. Gaumer、L. McDonald、N. Hodas、E. Jurrus，通过无监督形貌分析判别氧化铀的合成路径，Journal of Nuclear Materials（ 2021 152983doi:10.1016/j.）。 0022311521002063
jnucmat.2021.152983
URL https://linkinghub.elsevier.com/retrieve/pii/S[8]

C. A. Nizinski、C. Ly、C. Vachet、A. Hagen、T. Tasdizen、L. W. McDonald IV，铀矿石精矿形貌卷积神经网络预测的不确定性与模型泛化能力表征，Chemometrics and Intelligent Laboratory Systems 225（2022年）104556。[9] C. Ly、A. M. Olsen、I. J. Schwerdt、R. Porter、K. Sentz、L. W. McDonald、T. Tasdizen，量化 U3O8 形态特征的新方法：使用深度学习模型进行核取证，

用于核取证的 U3O8 形态特征，Journal of Nuclear Materials 517（2019年）128–137。[10] G. J. Sz´ekely、M. L. Rizzo，高维中的距离相关独立性 t 检验，Journal of Multivariate Analysis 117（2013年）193–213。[11] B. R
´ 3 2009 1295 1298 doi:10.1214/09-AOAS312F
emillard，《布朗距离协方差》讨论，Annals of Applied Statistics（）–。。[12] M. R. Kosorok，《布朗距离协方差》讨论，Annals of Applied Statistics
3 2009 1270 1278 doi: （）–。。10.1214/09-AOAS312B
[13] M. Heusel、H. Ramsauer、T. Unterthiner、B. Nessler、S. Hochreiter，采用双时间尺度更新规则训练的生成对抗网络收敛于局部纳什均衡，载于：I. Guyon、U. V. Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan、R. Garnett（编），《神经信息处理系统进展》，第30卷，Curran Associates, Inc.，2017年。URL https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf[14] A. Hagen、S. Jackson、J. Kahn、J. Strube、I. Haide、K. Pazdernik、C. Hainje，高维 Kolmogorov–Smirnov 距离的加速计算（6 2021年）。[15] A. Gretton、K. M. Borgwardt、M. J. Rasch、B. Sch¨olkopf、A. Smola，核双样本检验，《机器学习研究杂志》13（1）（2012年）723–773。[16] C. Hainje、C. Nizinski、S. Jackson、R. Clark、F. Heller、I. Schwerdt、E. Buck、D. Meier、A. Hagen，使用向量量化变分自编码器研究氧化钚的工艺历史与现象学，已投稿至 Chemometrics and Intelligent Laboratory Systems（12 2022年）。[17] A. Gretton、K. M. Borgwardt、M. Rasch、B. Sch¨olkopf、A. J. Smola，双样本问题的核方法，2006年。URL https://proceedings.neurips.cc/paper/2006/file/e9fb2eda3d9c55a0d89c98d6c54b5b3e-Paper.pdf[18] R. Fisher，《实验设计》，Hafner，1935年。[19] D. Lopez-Paz、M. Oquab，重新审视分类器双样本检验，第五届国际学习表征会议，ICLR 2017 - Conference Track Proceedings（2016年）。 doi:10.48550/arxiv.1610.06545。[20] E. B. Wilson，概率推断、继承律与统计推断，《美国统计学会杂志》22（1927年）209。 doi:10.2307/2276774。URL https://www.jstor.org/stable/2276774?origin=crossref

附录 A. 基于分类器的双样本检验

我们可以将基于分类器的双样本检验表述为从比例为 λ 的二项分布中抽取 N 次实现。对于唯一值的数量 N_v，获得正确预测的概率必须不高于随机机会，即比例 λ 必须等于 $\frac{1}{N_v}$。我们可以通过测量平衡准确率 a 来计算该比例的一次实现。因此，如果准确率与 $\frac{1}{N_v}$ 存在显著差异，则表明已预测到某种效应。利用这些事实，我们可以确定达到显著性所需的样本数量 N。首先，从二项比例的威尔逊近似置信区间 [20], 开始，我们可以写成

$$p \approx \frac{aN + \frac{1}{2}z^2}{N + z^2} \pm \frac{z}{N + z^2}\sqrt{a\left(1-a\right)N + \frac{z^2}{4}}, \tag{A.1}$$

其中 z 是与显著性水平 α 相对应的标准正态分布分位数（即对于 95% 而言为 $z = 1.96$ ）。当存在效应时，随机机会比例的置信区间上限 p_0 将小于分类器置信区间的下限 p_a。随后，我们可以求解 N，从而得到达到显著性的条件为

$$N = z^2 \frac{\left[-\left(N_v a\right)^2 + 2N_v^2 a - 2N_v a + 2N_v - 1\right]}{\left(N_v a\right)^2 - 2N_v a + 1}. \tag{A.2}$$