This phenomenon states that, when one attempts to improve and no need exists to incorporate many variables in a model. Nonetheless, one may argue that the so-called software packages that are well developed for classical and, in part, to the ready-to-use computational and statistical software packages that are well developed for classical techniques. Moreover, many scientists working in a specific field of high-dimensional statistical learning are either not aware of other existing machineries in the field or are not willing to try them out. The primary goal in this work is to bring together various machineries of high-dimensional analysis, give an overview of the important results, and present the operating conditions upon which they are grounded. When appropriate, readers are referred to relevant review articles for more information on a specific subject.

KEYWORDS: high-dimensional analysis, shrinkage, ridge estimation, sparsity, curse of dimensionality, G-analysis, double asymptotics, random matrix theory, Kolmogorov asymptotics

**SUPPLEMENT: Statistical Systems Theory in Cancer Modeling, Diagnosis, and Therapy**

**ABSTRACT:** High-dimensional data generally refer to data in which the number of variables is larger than the sample size. Analyzing such datasets poses great challenges for classical statistical learning because the finite-sample performance of methods developed within classical statistical learning does not live up to classical asymptotic premises in which the sample size unboundedly grows for a fixed dimensionality of observations. Much work has been done in developing mathematical-statistical techniques for analyzing high-dimensional data. Despite remarkable progress in this field, many practitioners still utilize classical methods for analyzing such datasets. This state of affairs can be attributed, in part, to a lack of knowledge and, in part, to the ready-to-use computational and statistical software packages that are well developed for classical techniques. Moreover, many scientists working in a specific field of high-dimensional statistical learning are either not aware of other existing machineries in the field or are not willing to try them out. The primary goal in this work is to bring together various machineries of high-dimensional analysis, give an overview of the important results, and present the operating conditions upon which they are grounded. When appropriate, readers are referred to relevant review articles for more information on a specific subject.

**COMPETING INTERESTS:** Author discloses no potential conflicts of interest.

**FINANCIAL DISCLOSURE:** This work was partially supported by the Nazarbayev University Social Policy Grant. The author confirms that the funder had no influence over the study design, content of the article, or selection of this journal.

**FUNDING:** This work was partially supported by the Nazarbayev University Social Policy Grant. The author confirms that the funder had no influence over the study design, content of the article, or selection of this journal.

**PEER REVIEW:** Four peer reviewers contributed to the peer review report. Reviewers’ reports totaled 424 words, excluding any confidential comments to the academic editor.

**Academic Editor:** J. T. Efird, Editor in Chief

**Type:** Review

**Citation:** Zollanvari. High-Dimensional Statistical Learning: Roots, Justifications, and Potential Machineries. Cancer Informatics 2015:14(5S) 109–121
doi: 10.4137/CIn.S30804

**Keywords:** high-dimensional analysis, shrinkage, ridge estimation, sparsity, curse of dimensionality, G-analysis, double asymptotics, random matrix theory, Kolmogorov asymptotics

---

**Introduction**

Classical statistical techniques have been fashioned for situations in which the number of data points is much larger than the number of variables.\(^1\) This is in large part due to the classical notion of statistical consistency, which guarantees the performance of a statistical technique in situations where the number of measurements unboundedly increases \((n \rightarrow \infty)\) for a fixed dimensionality \(p\) of observations.\(^2-5\)

However, even though many modern datasets are characterized by a number of variables far exceeding the sample size, many practitioners still utilize classical learning methods to extract information out of such datasets. This state of affairs can be attributed, in part, to a lack of knowledge and, in part, to the ready-to-use computational and statistical software packages that are well developed for classical techniques. Nonetheless, one may argue that the so-called curse of dimensionality phenomenon in statistical learning can serve as a justification for utilizing classical techniques, and no need exists to incorporate many variables in a model. This phenomenon states that, when one attempts to improve performance by increasing the number of variables for a given number of data points, the performance improves up to a certain point, after which it starts deteriorating.\(^6\) This phenomenon seems a justification for reducing the number of variables (dimensionality reduction) to a small number, perhaps much less than the sample size. In this reduced feature space, we are then “safe” to apply classical schemes because now the sample size is potentially much larger than the number of variables. The effect of the curse of dimensionality and its implications will be described in more detail later.

Using classification as an archetype, the dimensionality reduction generally follows a common methodology: 1) use a classification rule, including a feature selection, to design a classifier, and 2) use an error estimation rule to estimate the error of the designed classifier. The performance of many widely used classifiers is guaranteed in situations where \(n >> p\). They are designed to converge (in probability) to the Bayes classifier (optimal classifier) if \(n \rightarrow \infty\) and \(p\) is fixed. Likewise, the performance of many error estimation rules lives up to similar asymptotic premises. Therefore, the feature selection strategy serves as an interface in order to scale.
the complexity of data to one that can be studied through classical methods. Fortunately, two mathematical–statistical machineries exist that are specifically designed to serve in high-dimensional settings: 1) shrinkage, and 2) the Girko $G$-analysis. These frameworks can serve as potential machineries in order to develop mathematical models suitable for analysis in situations wherein the dimensionality of observations is comparable or potentially larger than the sample size. While the shrinkage estimation is grounded on the sparsity principle, $G$-analysis, in its simplest form, is based on double asymptotics $n \to \infty$, $p \to \infty$, $p/n \to c$, $0 < c < \infty$, as well as on some conditions on the existence of moments of random variables involved.\(^7\) However, $G$-analysis makes no assumption on the sparsity of the parameters to be estimated. Note that having the last two conditions, that is, $p \to \infty$ and $p/n \to c$, implies first $n \to \infty$.

The sparsity principle imposes an assumption on the nature of the probabilistic structure of observations; it assumes that only a small number of predictors contribute to the response.\(^8\) In other words, while the curse of dimensionality restricts the number of variables feeding a model (by a subset selection strategy), the sparsity principle, on the other hand, does not restrict the number of variables. Instead, a model is potentially trained on all variables, and it has a good performance if the parameter space is sparse. While the effect of parameter sparsity on the behavior of shrinkage estimation has been studied to some extent, the effect of the curse of dimensionality on $G$-analysis has been generally left unexplored. Understating the effect of the peaking phenomenon or the curse of dimensionality is important because, if it can be avoided, then we can see the $G$-analysis as a potential machinery to follow in situations where the parameter sparsity is not well justified. It might be argued that there is nothing wrong with the classical methodologies (which work well when $n \to \infty$ and $p$ is fixed) because (in the context of classification) ultimately it is the error of the designed classifier that matters, and, if classical methodology does not work, then the price paid will be poor performance. This is a legitimate argument as long as the cost is negligible. Unfortunately, this is not always the case, as the next paragraph illustrates.

Let us consider genomic datasets as a prototypical example of a modern, high-dimensional, small-sample dataset. In 2005, Michiels et al.\(^9\) challenged the validity and repeatability of several microarray-based research studies. They reported that a reanalysis of data from the seven largest published microarray-based studies, which attempted to predict the prognosis of cancer patients, revealed that five of those seven did not really classify patients better than a random assignment. There were other studies aimed at reproducing the published results of such prognosis studies, but they too generally failed\(^10,11\). The consequence of the failures in many genomic research studies has been brought into sharp focus by Dr. J. Woodcock, Director of the Center for Drug Evaluation and Research (CDER) at the U.S. Food and Drug Administration. She stated, “We may be out of the general skepticism phase, but we are in the long slog phase”\(^12\). In listing barriers to “coming up with the right diagnostics,” she estimated that 75% of published biomarker applications are not replicable: “This poses a huge challenge for industry in biomarker identification and diagnostics development”\(^13\).

From a technical point of view, the irreproducibility crisis of the results that we are facing today\(^14,15\) can be attributed in large part to the nature or misuse of our classical statistical techniques.\(^16,17\) This state of affairs could have been preventable years ago if we took the following lines noted by Ronald A. Fisher, one of the first biologist–geneticist–statisticians, more seriously. In 1925, he said\(^18,19\):

Little experience is sufficient to show that the traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow! The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their metrics does it seem possible to apply accurate tests to practical data.

**Sparsity and Shrinkage Estimation**

Let $x$ be a realization of a random vector of $p$ dimension that is normally distributed with unknown mean $\theta$ and identity covariance matrix, ie, $x \sim \mathcal{N}(\theta, I)$. Our goal is to estimate the vector $\theta$. One may consider $x$ as being the sample mean constructed from $n$ observations, in which case the covariance matrix is $I/n$; however, as in many other studies,\(^20\) we consider the more convenient canonical form that is obtained after proper scale transformations, which results in the identity covariance matrix $I$.

In a seminal work,\(^20\) Stein astonished the statistical community by showing that, if we consider the total squares of the errors as the loss function and $p \geq 3$, then there exists a class of estimators of $\theta$ that uniformly has a smaller risk than that of the regular maximum likelihood estimator. In other words, the “usual estimator” $\hat{\delta}_m(x) = x$ is inadmissible for $p \geq 3$. Later, James and Stein gave an explicit form of an estimator that uniformly dominates $\hat{\delta}_m(x)$,\(^21\) ie, uniformly has a smaller risk than $\hat{\delta}_m(x)$; for an estimator $\delta$, the risk $R(\theta, \delta)$ is the expectation of the loss function over the sample space, given by

$$R(\theta, \delta) = \mathbb{E}([|\theta - \delta|]^2),$$

(1)

where $||y||^2 = y^Ty$ is the $\ell_2$ norm.

This achievement led to a large body of work on many aspects of the problem, proposing estimators for situations in which the covariance matrix of the normal population is known or unknown, extending this estimation procedure.
to non-normal populations, the Bayesian justification of the James–Stein estimator, and various attempts to improve upon the James–Stein estimator. Here, it is not possible to summarize the large amount of work done in this direction. We describe some of the key developments in the field, but for more information the readers are referred to other works.\textsuperscript{24–28}

The James–Stein estimator is given by\textsuperscript{31}

$$\hat{\delta}_p = \left(1 - \frac{p - 2}{\|x\|^2}\right) x.$$  \hfill (2)

It turns out that $\hat{\delta}_p$ is itself inadmissible and has a peculiar behavior for small values of $\|x\|$. In particular, the shrinkage factor \(\left(1 - \frac{p - 2}{\|x\|^2}\right)\) becomes negative, in which case the sign of each component of the vector $x$ changes. Baranchik proposed an improved estimator that dominates $x$.\textsuperscript{29,30} This estimator is obtained by taking the “positive part” of the James–Stein estimator as

$$\hat{\delta}_p^+ = \left(1 - \frac{p - 2}{\|x\|^2}\right)^+ x,$$  \hfill (3)

where for a real number $k, k^+ = \max(0,k)$. The estimator $\hat{\delta}_p^+$ has also an undesirable property that, when $\|x\|^2 \leq \frac{p}{p - 2}$, each element of the vector $\theta$ is estimated by zero. From the well-known results on smoothness of admissible estimators,\textsuperscript{31} it follows that $\hat{\delta}_p$ itself is inadmissible. Baranchik\textsuperscript{30} widened the class of Stein’s estimator by showing that any estimator of the form

$$\hat{\delta}_p = \left(1 - \frac{\phi(x)}{\|x\|^2}\right) x,$$  \hfill (4)

where $\phi(.)$ is (i) monotone, nondecreasing, and (ii) $0 \leq \phi(.) \leq \frac{p - 2}{\delta_{m}^2}$, dominates the usual estimator. Note that $\hat{\delta}_p$ is a special case of $\hat{\delta}_p$ for $\phi(.)$ any constant satisfying (ii). Several other estimators have been proposed,\textsuperscript{32,33} but Kubokawa’s estimator\textsuperscript{14} is an estimator that dominates $\hat{\delta}_p$ and is admissible. A compelling intuition behind the James–Stein estimator and all previous extensions is as follows: Suppose an estimation of $\theta$ is desired and, rather than using the usual estimator $x$, we use an estimator $\hat{\delta}_p = \alpha x$, where $\alpha$ is determined by minimizing the risk function. The risk of this estimator is given by

$$R(\theta, \hat{\delta}_p) = \mathbb{E}\left[\|\theta - \hat{\delta}_p\|^2\right] = \left\|1 - \alpha\right\|^2 + p\alpha^2.$$

One can verify that the optimal choice of $\alpha$ that minimizes $R(\theta, \hat{\delta}_p)$ is given by

$$\alpha_{opt} = 1 - \frac{\rho}{\|\theta\|^2 + \rho},$$  \hfill (6)

which results in the minimum risk (the Bayes risk), given by

$$R(\theta, \hat{\delta}_{opt}) = \frac{\rho \|\theta\|^2}{\|\theta\|^2 + \rho}.$$  \hfill (7)

However, note that with the choice of $\alpha$ in (6), $\alpha x$ is not an estimator anymore because it depends on the unknown $\|\theta\|$. However, since $E[\|x\|^2] = \|\theta\|^2 + \rho$, one can estimate $\|\theta\|^2 + \rho$ in the denominator of (6) by $\|x\|^2$ and obtain the estimator $(1 - \rho/\|x\|^2)x$, which is in the form of the James–Stein estimator. An even better estimator is obtained if we replace $\rho$ by $\rho - 2$, which results in $\hat{\delta}_p$.\textsuperscript{23,24} Since the risk of the “usual estimator” $\hat{\delta}_{m}(x)$ is $R(\theta, \hat{\delta}_{m}) = \rho$, we have

$$R(\theta, \hat{\delta}_p) - R(\theta, \hat{\delta}_{m}) = \frac{\rho}{\rho + \|\theta\|^2}.$$  \hfill (8)

It is evident that, when $\|\theta\|$ is large (compared to $\rho$), then $R(\theta, \hat{\delta}_p) = R(\theta, \hat{\delta}_{m})$ and we do not gain much by shrinking. However, for small values of $\|\theta\|, R(\theta, \hat{\delta}_p) = 1 \ll \rho = R(\theta, \hat{\delta}_{m})$. This shows that for a model (here the mean of a multivariate Gaussian distribution) that is sparse (contains many zero elements), the larger the dimension, the more we gain by shrinking.

Similar to the James–Stein estimation, ridge estimation is a type of shrinkage originally proposed in\textsuperscript{35} and developed further by many researchers. Consider the linear model

$$y = X\beta + \varepsilon,$$  \hfill (9)

where $y$ is an $n$-dimensional observation vector, $X$ is a known $n \times p$ matrix, $\beta = [\beta_1, \beta_2, \ldots, \beta_p]^T$ is a $p$-dimensional parameter vector to be estimated, and $\varepsilon$ is an $n$-dimensional error vector with mean 0 and covariance matrix $\sigma^2 I_n$. If we assume $X$ is a full (column) rank matrix ($p < n$), the ordinary least-square solution to this familiar linear model is given by

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$  \hfill (10)

However, when $p > n$, the solution (10) does not exist because $X^T X$ becomes degenerate. Even the solution obtained by generalized inverse form of matrix $X^T X$ does not work well. A possible solution was proposed by Hoerl and Kennard,\textsuperscript{35–37} who replaced the residual sum of squares by its $\ell_2$ penalized form, given by

$$L_2(\beta) = \|y - X\beta\|^2 + \lambda \|\beta\|^2,$$  \hfill (11)

where $\lambda > 0$ denotes a penalty factor controlling the length of $\beta$. Minimizing $L_2(\beta)$ results in the so-called ridge regression, given by

$$\hat{\beta} = (X^T X + \lambda I_p)^{-1} X^T y.$$  \hfill (12)

In this way, the inverse of possibly ill-conditioned $X^T X$ is stabilized by adding the scalar matrix $\lambda I_p$. The value of $\lambda$
is assumed to be a meta-parameter of the procedure and, in general, is estimated from a model-selection strategy such as cross-validation. A family of estimators closely related to the ridge estimator is obtained by considering the $\ell_1$ penalty factor in (11). In that case, the goal is to minimize

$$L_1(\beta) \triangleq \|y - X\beta\|^2 + \lambda \|\beta\|.$$  (13)

Here, $\lambda$ determines a trade-off between the approximation error ($\|y - X\beta\|^2$) and the sparsity of parameters $\|\beta\|$. Of course, a more natural way for such a trade-off would have been the $\ell_0$ penalty factor ($\sum_{i=1}^p 1(\beta_i = 0)$), but replacing $\ell_1$ with $\ell_0$ in (13) and trying to minimize the resulting risk function would be an NP-hard problem. The "least absolute shrinkage and selection operator" (lasso) technique proposed by Tibshirani is used to minimize $L_1(\beta)$. Tibshirani formulates the following quadratic programming problem:

$$\begin{align*}
\text{minimize} & \quad \|y - X\beta\|^2 \\
\text{subject to} & \quad \|\beta\| \leq t
\end{align*}$$  (14)

which is equivalent to minimizing $L_1(\beta)$; to wit, for any $\lambda$ in $L_1(\beta)$, there exists a $t \geq 0$ such that (14) has the same solution as minimizing $L_1(\beta)$. The key element in the popularity of the lasso is its ability to perform estimation and model selection at the same time. For a design matrix $X$ such that $X^T X = I_p$ (orthonormal), understanding the mechanism of the lasso’s model selection is straightforward. In this case, the lasso estimate of $\beta_i$, denoted by $\hat{\beta}_i$, is obtained by the soft shrinkage operator originally proposed by Donoho and Johnson: to wit, we have

$$\hat{\beta}_i = \text{sign}(\beta_i)\left(|\beta_i| - \lambda\right)^+, \quad (15)$$

where $\beta_i^+$ is the solution of ordinary least-square and $(x)^+ = \max(x, 0)$. Evidently, for $\lambda > |\beta_i|$, $\hat{\beta}_i$ is exactly zero. Therefore, in this case, having $\lambda > |\beta_i|$, $i = 1, 2, \ldots, k$, is equivalent to choosing a subset of size $k$ variables and setting all other parameters to 0. Therefore, for a large value of $\lambda$, many $\beta_i$ values are set to 0 and the solution becomes sparse. For a general design matrix $X$, minimizing $L_1(\beta)$ has no closed-form solution. Solving (14) using the algorithm proposed by Tibshirani is not efficient for large values of $p$, but efficient algorithms have emerged to compute the solution. There has been a large body of work on generalizing the idea of the lasso, eg, elastic net, adaptive lasso, fused lasso, and graphical lasso. The readers are referred to for a review of the various generalizations of the lasso technique. Before the emergence of the idea of the lasso, the idea of using $\ell_1$ norm for sparse representation of signals was used in the signal processing community (also see Appendix I in Ref. 39). Nevertheless, in signal processing, the idea was generalized and formalized later in Ref. 51, resulting in the basis pursuit (BP) framework. In signal processing applications, the BP is used to decompose a signal ($y$) into an optimal superposition of discrete dictionary waveforms ($X$) such as Fourier dictionary, wavelet dictionary, etc. The optimality is based on minimizing the $\ell_1$ norm of coefficients $\beta$ of the representation such that the solution becomes an affine subspace in $\mathbb{R}^p$. Formally, BP solves

$$\begin{align*}
\text{minimize} & \quad \|\beta\| \\
\text{subject to} & \quad X\beta = y
\end{align*}$$  (16)

Therefore, by minimizing the $\ell_1$-norm of coefficients, we are seeking for the sparsest possible solution. In the presence of noise, BP is used by solving a quadratically constrained linear program, which is a trade-off between a quadratic misfit and the $\ell_1$ norm of coefficients:

$$\begin{align*}
\text{minimize} & \quad \|\beta\| \\
\text{subject to} & \quad \|y - X\beta\|^2 \leq \varepsilon
\end{align*}$$  (17)

in which $\varepsilon > 0$ controls the amount of trade-off. As we have seen, the James–Stein estimation and $\ell_1$ shrinkage (lasso-based or BP-based methods) are well suited to situations where the underlying model is sparse. Next, we review independent machinery for high-dimensional analysis in which the assumption of sparsity does not play a role.

**Curse of Dimensionality and G-analysis**

Generalized consistent estimation (also known as Girko G-analysis) is a technique to construct estimators specific to situations in which the dimension is comparable to the number of samples. In this setting, an estimator is constructed such that it converges to the actual parameter in a “double-asymptotic” sense, to wit, in an asymptotic scenario in which dimension and sample size increase in a proportional manner, eg, $n \to \infty, p \to \infty$, and $p/n \to \varepsilon > 0$. In this framework, the sparsity principle does not play a role. However, if the curse of dimensionality is intrinsic to frequentist statistics, then regardless of the model we use, there will still be a large gap between the complexity that we can capture through our model and the complexity of the phenomenon under study (if, of course, the phenomenon is complex per se). Therefore, in the subsequent discussion, we first examine the curse of dimensionality, its origin, and implications.

**Curse of dimensionality.** The curse of dimensionality, also known as the “peaking phenomenon” or the “Hughes phenomenon”, is generally considered as the known fact accounting for dimensionality reduction and feature selection. Regarding the peaking phenomenon, McLachlan stated:

---

**Zollanvari**

112 | CANCER INFORMATICS 2015:14(S5)
For training samples of finite size, the performance of a given discriminant rule in a frequentist framework does not keep on improving as the number \( p \) of feature variables is increased. Rather, its overall unconditional error rate will stop decreasing and start to increase as \( p \) is increased beyond a certain threshold, depending on the particular situation.

Jain and Waller stated\(^{52} \):

Thus, even if the cost of taking measurements is negligible, there exists an optimum measurement complexity, which is a function of the number of available training samples and the probability structure of the model.

Chandrasekaran and Jain pointed out:

It is known that, in general, the number of measurements in a pattern classification problem cannot be increased arbitrarily, when the class-conditional densities are not completely known and only a finite number of learning samples are available. Above a certain number of measurements, the performance starts deteriorating instead of improving steadily.

See Ref. 55 or p. 561 in Ref. 56 for more comments about this phenomenon. The first observation of peaking phenomenon is attributed to the work of Hughes.\(^{53} \) However, the peaking observed by Hughes was shocking to many scientists since it was contrary to the previously reported results on the lack of peaking for Bayes (optimal) classifiers. Hughes noted, “If insufficient sample data are available to estimate the pattern probabilities accurately, then a Bayes recognizer is not necessarily optimal.”\(^{53,57} \) Various researchers correctly criticized Hughes’ work by pointing out that the paradoxical peaking phenomenon observed therein was not real and was due to the estimate of unknown cell probabilities from the data.\(^{17-46} \) In other words, the peaking phenomenon observed by Hughes was essentially within a frequentist framework, not a Bayesian.

Nevertheless, it is now the general consensus that in the frequentist framework, the performance of a constructed classifier does not keep improving as more features are added. To be more precise, it is assumed that there is a certain point after which we should not keep adding features because the expected error rate of the classifier starts to increase (see above quotes as well as Refs. 6 and 56). Commonly, this certain point is referred to as the “optimal number of features”\(^{52,55} \). Nevertheless, all the aforementioned studies, and even terminologies such as curse of dimensionally or peaking phenomenon, give the impression that we should not learn from a large number of variables when a finite (and perhaps relatively small) number of samples is available.

Here, I shall try to convince you that the curse of dimensionality is not a phenomenon intrinsic to the frequentist framework. Instead, it is an artifact of many contemporary frequentist approaches. However, let us first review a few theoretical works that show peaking phenomenon in a frequentist setting.

In Ref. 6, the authors studied the peaking phenomenon in the context of discrete classification using a histogram rule. They considered multinomial distributions governing the data and characterized the expected error rate of the histogram rule over both sample space and a uniform prior distribution on the multinomial parameters. However, the complexity of the expression obtained there for the expected error rate did not let them achieve an analytical solution for the optimal dimension or analytical proof for the existence of a dimension at which the expected error rate is minimized.

Another work in this context is that of Jain and Waller,\(^ {52} \) who analytically studied the peaking phenomenon in connection with linear discriminant analysis (LDA) and in the context of Gaussian multivariate models. They used Bowker and Sitgreaves’s\(^ {61,62} \) approximation of the expected error of LDA to determine an expression for the minimal increase in \( \delta^2 \) that justifies adding another feature (“avoid peaking”), where \( \delta^2_p \) is the Mahalanobis distance given by

\[
\delta^2_p = (\theta_i - \theta_j)^T \Sigma^{-1} (\theta_i - \theta_j),
\]

where \( p \) denotes the dimensionality. They showed that the minimal increase in \( \delta^2_p \) (assuming \( \delta^2_p \gg \frac{1}{p} \)) to keep the same expected error rate is given by

\[
\delta^2_{p+1} - \delta^2_p = \frac{\delta^2_p}{n - 3 - p},
\]

where \( n \) in the total number of samples in both classes. They also used this minimal increase expression to obtain various curves of expected error rate to show the peaking phenomenon. Nevertheless, their results were specific to LDA. Their observation should not be generalized to other models or even interpreted as a proof of omnipresence of peaking phenomenon in the frequentist framework (at least the way it is commonly presented). Moreover, the asymptotic expansion that Bowker and Sitgreaves used to develop their approximation of the expected error was essentially developed under the classical \( n \to \infty \) and fixed \( p \) regime, which has a poor performance in a high-dimensional setting.

The salient point is that, even with the existing classifiers that have been developed through the classical statistical framework \( (n \to \infty, p \text{ fixed}) \), the peaking phenomenon is not what is commonly perceived. To show this, in the following, we present an example in which even after the so-called optimal number of features has been found, we still keep adding features to the model. From earlier discussion, recall that the optimal number of features is generally considered as the number after which adding more features deteriorates the performance. However, in this example, we observe that after
CanCer Informatics

n -
is the Mahalanobis distance given by (18) and (22) are the sample means \( \mathbf{x}_i \) of \( \mathcal{N}(\mu_i, \Sigma) \) for \( i = 0, 1 \), where \( \mu_0 \) and \( \Sigma \) denote the mean and the covariance matrix, respectively. Consider the following discriminant function:

\[
W(\mathbf{x}_0, \mathbf{x}_1) = \left( \mathbf{x} - \frac{1}{n_0} \sum_{j=1}^{n_0} \mathbf{x}_j \right)^T \left( \frac{1}{n_1} \sum_{j=n_0+1}^{n_0+n_1} \mathbf{x}_j - \frac{1}{n_1} \sum_{j=n_0+1}^{n_0+n_1} \mathbf{x}_j \right),
\]

where \( \mathbf{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_j \) and \( \mathbf{x}_i = \frac{1}{n_i} \sum_{j=n_i+1}^{n_0+n_i} \mathbf{x}_j \) are the sample means for each class. The Euclidean-distance classifier (EDC) is given by

\[
\psi(\mathbf{x}) = \begin{cases} 1, & \text{if } W(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}) \leq 0 \\ 0, & \text{if } W(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}) > 0 \end{cases},
\]

That is, the sign of \( W(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}) \) determines the classification of the sample point \( \mathbf{x} \). The true error of \( \psi(\mathbf{x}) \), denoted by \( \epsilon \), is defined to be the probability of misclassification:

\[
\epsilon_{n,p} = \alpha_0 \epsilon_0 + \alpha_1 \epsilon_1,
\]

where \( \alpha_i \) is the prior probability for class \( i \) and \( \epsilon_i \) is the error contributed by class \( i \), which is given by

\[
\epsilon_i = P \left( \mathbf{x} \right) W(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}) \leq 0 \mid \mathbf{x} \in \Pi_i, \mathbf{x}_0, \mathbf{x}_1 \right).
\]

From (23), we have

\[
E_i = \Phi \left( \frac{(-1)^{\epsilon_i} \left( \mathbf{x}_0 - \mathbf{x}_1 \right)^T \left( \mathbf{x} - \mathbf{x}_i \right)}{\sqrt{\left( \mathbf{x}_0 - \mathbf{x}_1 \right)^T \Sigma \left( \mathbf{x}_0 - \mathbf{x}_1 \right)}} \right).
\]

Assuming \( n_0 = n = n_0, \alpha_0 = \alpha_1 \), and \( \Sigma = \sigma^2 \mathbf{I} \), from the results of Raudys \( 64,65 \) (also see Refs. 66 and 67) we can approximate the expected error (over the sampling distribution) of EDC by

\[
E \left[ \epsilon_{n,p} \right] = \Phi \left( \frac{-\delta^2}{\sqrt{\delta^2 + 2f}} \right),
\]

where \( \delta^2 \) is the Mahalanobis distance given by (18) and \( f = p/n \). This formula is also credited to A.N. Kolmogorov (see p. 3 of Ref. 68). Raudys’ approximations of finite sample expectation of true error for LDA and its derivatives (including EDC) are based on the basic form of the double-asymptotic approach \( (n \to \infty, p \to \infty, p/n \to \delta) \). In Ref. 69, the authors compared Raudys’ approximation to several well-known approximations that have been obtained by classical approaches \( (n \to \infty, p \text{ fixed}) \) and showed that the double-asymptotic approximations are significantly more accurate than classical approximations in analyzing the expected true error of LDA. In particular, even with \( n/p < 3 \), the double-asymptotic approximations yield excellent approximations that are far more accurate than others.\(^6\)

Let \( \mathbf{b}_i \) denote a column vector of size \( p \) with identical elements \( b \). In order to examine the peaking phenomenon, we consider two 1700-dimensional Gaussian distributions where \( \theta_0 = -\mathbf{1}, \theta_1 = \left[ 0.2^T(10), 0.05^T(190), 0.03^T(1200), 0^T(300) \right]^T \), and \( \Sigma = \mathbf{I}_{1700} \) with \( \mathbf{I}_p \) being the identity matrix of size \( p \). As described in Ref. 55, “best” features are generally added first and less useful features are added later. In our Gaussian model, the feature discriminative power is the same or reduces as we move to higher dimensions (see \( \theta_1 \)). Therefore, the first \( p \) features are our best features. At each \( p \), we train a classifier on 100 \( p \)-dimensional observations taken from the two aforementioned Gaussian distributions. We determine the expected error rate of EDC from (25). Figure 1 shows \( E[\epsilon_{n,p}] \) versus \( p \) in a logarithmic scale. The solid curve is obtained from (25). In order to ensure that this curve is not an artifact of using approximation (25), \( E[\epsilon_{n,p}] \) has been estimated by Monte Carlo simulations for a few dimensions as follows:

**The Monte Carlo simulation protocol.**

- **Step I**: Fix a pair of \( p \)-dimensional Gaussian distribution \( \Pi_0 \) and \( \Pi_1 \) with identity covariance matrices and means being the first \( p \) elements of vector \( \theta_0 \) and \( \theta_1 \), respectively. In simulations, we only consider \( p = 10, 65, 200, 535, \) and 1400.

\[
\begin{array}{c|c|c}
\text{Expected Error} & \text{Analytical expression} & \text{Monte carlo simulation} \\
\hline
0.40 & \text{---} & \text{---} \\
0.30 & \text{---} & \text{---} \\
0.20 & \text{---} & \text{---} \\
0.10 & \text{---} & \text{---} \\
0.05 & \text{---} & \text{---} \\
0.02 & \text{---} & \text{---} \\
\end{array}
\]

**Figure 1.** Expected error of Euclidean-distance classifier versus dimension for \( n_0 = n_1 = 100 \). The solid curve is obtained from theoretical results. The small circles are the result of simulation experiments for \( p = 10, 65, 200, 535, \) and 1400.
• Step II: From each distribution, generate a training set of size $n = 100$.
• Step III: Using the training sample, construct EDC using (20) and (21).
• Step IV: Find the true error of the constructed classifier using (22) and (24); this is possible because we have parameters of our model.
• Step V: Repeat Steps II–IV 500 times and take the average. The result is an estimate of $E[E_{n,p}]$.

The result of the Monte Carlo simulation for the five dimensions that we have considered is depicted by small circles in Figure 1: they align well with the theoretical results represented by the curve. As we see in this figure, as soon as we start adding more than 10 features to the EDC model, the performance starts deteriorating, but if we keep adding more and more features, at about $p = 65$, the performance again starts to improve. At $p = 200$, it has a local minimum and this behavior repeats one more time, resulting in a multi-hump curve. Interestingly, by considering all 1,700 features in the classifier, we obtain a performance better than the first local minimum at $p = 10$; to wit, $E[E_{100,1,700}] = 0.273 < 0.276 = E[E_{100,10}]$. Nevertheless, the best performance happens when $p = 1,400$ ($E[E_{100,1400}] = 0.254$). Note that the EDC model is essentially a variant of the LDA classifier, which, under a Gaussian model, converges to the Bayes classifier as $n \to \infty$ and $p$ fixed. It is natural to expect development of better classifiers from a mechanism such as the G-analysis framework, which is specifically designed for high-dimensional analysis. The conclusion to be drawn from this example is not to reject the peaking phenomenon – we can cite many examples that demonstrate that the peaking phenomenon is observed in the same way that is classically stated. Instead, this example demonstrates the following: 1) the way the curse of dimensionality is generally stated does not reflect what this phenomenon really is and may give a wrong impression to many practitioners, and 2) a compromise between complexity of the learning model and the number of predictors may achieve a better performance in a large-dimensional space than in a small one.

**Double asymptotics and G-analysis. An example from random matrix theory.** Random matrix theory (RMT) is a type of double-asymptotic analysis that is more focused on the analysis of spectral distribution of random matrices. The spectral distribution of random matrices is an important subject in multivariate analysis, as many statistics can be represented in terms of functionals of the spectral distribution of some matrices. Let $[x_{i,j}, i,j = 1,2,...]$ be a double array of i.i.d. random variables with mean zero and variance 1. Let $x_j \triangleq [x_{1,j}, x_{2,j},...,x_{N,j}]^T$ and define the data matrix $X$ such that $X = [x_1, x_2,..,x_N]$. The $p \times p$ sample covariance matrix is then defined by

$$C_{n,p} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T.$$  

The empirical spectral distribution (ESD) of matrix $C_{n,p}$ is given by

$$F_{C_{n,p}}(x) = \frac{1}{p} \sum_{i=1}^{p} 1_{[\lambda_i(C_{n,p}) \leq x]},$$  

where $\lambda_i(C_{n,p})$, $i = 1, 2, ..., p$ are the eigenvalues of $C_{n,p}$ and $1_{\{x\}}$ is the indicator function. We consider two cases: 1) $x_{i,j}$ has a standard normal distribution; and 2) $x_{i,j}$ is taken from $\{-1,+1\}$ with equal probability. A fundamental problem in high-dimensional analysis is to study the convergence of sequence of $F_{C_{n,p}}(x)$. The histograms in Figure 2A–C represent the empirical spectral distribution of one realization of matrix $C_{n,p}$ for three cases in which $p/n = 1/2$ and, (a) $p = 1,000$ and $n = 2,000$; (b) $p = 100$ and $n = 200$; and (c) $p = 20$ and $n = 40$. Under a double-asymptotic regime where $p \to \infty$, $n \to \infty$, $p/n \to c > 0$, the empirical distribution converges almost surely to a nonrandom distribution function $F_c$ with the density given by

$$f_c(x) = (1 - 1/c)^+ \delta(x) + \frac{1}{2\pi c} \sqrt{(x-a)^- (b-x)^+},$$

where $a = (1 - \sqrt{c})^2$, $b = (1 + \sqrt{c})^2$ and $\delta(x)$ is the delta function. This result was obtained by Marcenko and Pastur in 1967 and today is referred to as the M–P law. As we see in these histograms, for relatively large values of $p$ and $n$ [Figs. 2(A), (B)], there is a substantial agreement between the M–P law and the ESD of one realization of matrix $C_{n,p}$.

Figure 2D–2F show the result of comparison between the average empirical spectral distribution of $N$ realizations of covariance matrix $C_{n,p}$ with the limiting spectral distribution in which $N$ is 10, 100, and 10,000, for $p = 1,000, 100,$ and 20, respectively (Fig. 2/F). As we see, there is a substantial agreement between the average empirical distribution and the M–P law in all three cases. Convergence of histograms in Figures 2A and 2D to the same limiting density presents an interesting property of the double-asymptotic regime; to wit, in this asymptotic regime the dependencies of the results on the realization of $C_{n,p}$ disappear. Furthermore, the result is independent of the distribution of the elements in the double array: for both normal and binary random variables the limiting spectral distributions are identical.

The convergence of histograms in Figure 2(D)–(F) to the same density shows another interesting property of this operating regime: if we consider the average (expected) behavior of the ESD, the result of double asymptotics, although theoretically valid for $p \to \infty$ and $n \to \infty$, agrees well with the empirical result even for situations in which $p$ and $n$ are relatively small (Fig. 2F). Note that in many practical situations we are interested in average behavior of a statistic and, in this regard, double asymptotics can serve as a potential machinery for analyzing and synthesizing statistics. It is illuminating to compare the result of double-asymptotic analysis with the classical statistical learning.
asymptotic analysis. The only information that we have from the classical asymptotic analysis is that, as \( n \to \infty \) and for fixed \( p \), the distribution of \( \lambda(C_p^n) \) converges (almost surely) to a delta function at 1. Clearly, this information is completely useless in estimating the empirical spectral distribution in all cases considered in this example.

**Emergence of double-asymptotic analysis.** In the past few decades, double-asymptotic analysis, in general, and RMT, in particular, have found eminent roles in various disciplines, including nuclear physics, statistical mechanics, signal processing, wireless communications, biology, and economics.\(^{70,72}\) Some scientists, such as Raj Nadakuditi,\(^{73}\) believe that RMT is “somehow buried deep in the heart of nature”. The first account of double-asymptotic analysis can be traced back to studying the limiting spectral distribution of random matrices of large dimension. This analysis was done by Eugene P. Wigner, the Nobel Prize winning physicist who, in the context of quantum physics and in connection with the energy levels of heavy nuclei, proved that the expected spectral distribution of Wigner matrices of increasing dimension converges to the semicircle law.\(^{74,76}\) In quantum physics, any measurable physical quantity (a dynamical variable) of a system is represented by a self-adjoint operator (commonly referred to as the Hamiltonian) that acts in the state space. The Hamiltonian operator acting in state space can be represented in the matrix representation, resulting in matrix mechanics – which Heisenberg used to formulate quantum mechanics in the first place.\(^{77,78}\) Concerning the rise of RMT in physics, Freeman Dyson writes\(^{79}\):

> By assuming all states of a very large ensemble to be equally probable, one obtains useful information about the overall behavior of a complex system, when the observation of the state of the system in all its detail is impossible. What is here required is a new kind of statistical mechanics, in which we renounce exact knowledge not of the state of a system but of the nature of the system itself. We picture a complex nucleus as a ‘black box’ in which a large number of particles are interacting according to unknown laws. The problem then is to define in a mathematically precise way an ensemble of systems in which all possible laws of interaction are equally probable.

And concerning the randomness of the Hamiltonian that appears in Schrödinger’s equation, Mehta writes,\(^{80}\)

> In the case of the nucleus, however, there are two difficulties. First, we do not know the Hamiltonian and, second, even if we did, it would be far too complicated to attempt to

---

**Figure 2.** (A)–(C) Comparing the empirical spectral distribution of one realization of the covariance matrix with the limiting spectral distribution: (A) \( p = 1,000, \ n = 2,000; \) (B) \( p = 100, \ n = 200; \) (C) \( p = 20, \ n = 40. \) (D), (E) Comparing the average empirical spectral distribution of \( N \) realizations of the covariance matrix with the limiting spectral distribution: (D) \( p = 1,000, \ n = 2,000, \) and \( N = 10; \) (E) \( p = 100, \ n = 200, \) and \( N = 100; \) (F) \( p = 20, \ n = 40, \) and \( N = 10,000. \) Marčenko and Pastur\(^{71}\) obtained the closed-form solution of the limiting spectral distribution by using the double-asymptotic framework.
solve the corresponding equation. Therefore, from the very beginning we shall be making statistical hypotheses on \( \mathbf{H} \) [Hamiltonian], compatible with the general symmetry properties. Choosing a complete set of functions as a basis, we represent the Hamiltonian operators \( \mathbf{H} \) as matrices. The elements of these matrices are random variables whose distributions are restricted only by the general symmetry properties we might impose on the ensemble of operators.

Here Mehta eloquently describes the reason why pioneers such as Wigner and Dyson associated the randomness of the Hamiltonian and, consequently, its corresponding matrix representation with the complexity of the nucleus (see Ref. 79 for more details). Beside the curse of dimensionality that we discussed earlier, the principle of parsimony is another motivation for an immediate use of dimensionality reduction regardless of the complexity of the phenomenon under study. While the principle of parsimony tells us that a simpler model is preferable to a competing complex model, it does not tell that all phenomena are simple. While the sparsity of parameters describing a phenomenon seems a tempting idea, not all phenomena are sparse. A heavy nucleus is a perfect example of a complex system. What Wigner did was not to create a parsimonious model to describe such a system. Instead, in his groundbreaking work, he increased the complexity of model to infinity by considering random matrices of infinite dimension.\(^{64,75}\) As described in Ref. 81, the Bayesian philosophy dominated the statistics in the nineteenth century, but twentieth century was more a frequented one. I believe that after the long 250-year-old debate between Bayesians and frequentists, if there is a revolution in statistical learning community (if not yet), it happens in shifting the low-dimensional analytical paradigm to a high-dimensional one (see the R.A. Fisher’s quote in the introduction).

In the last 60 years, there have been an enormous number of studies in the context of random matrices of increasing dimension. The field has been developed to a large extent in the hands of F. J. Dyson,\(^{79,82,83}\) M. L. Mehta,\(^{84–87}\) L. A. Pastur,\(^{71,88,89}\) V. L. Girko,\(^{91–96}\) Z. D. Bai,\(^{97–100}\) and Y. Q. Yin.\(^{101–104}\) As estimated in Ref. 70, there has been more than 2,500 publications in the field from 1955 to 2004. The readers are encouraged to consult Refs. 70,72,80 for historical surveys of some of important results in the field.

A set of independent, but closely related work to previous studies, is the work of Raudys, Deev, Meshalkin, Serdobolskii, and Fujikoshi on the application of double asymptotics in classification.\(^{66,67,105–112}\) This body of work is formalized as follows: consider a sequence of Gaussian discrimination problems with a sequence of parameters and sample sizes: \( \left( \mathbf{\mu}_p, \mathbf{\Sigma}_p, n_p, p \right)_{p=1,2,...}, \) where the means and the covariance matrix are arbitrary. Let us denote the limit under

\[
\lim_{n \to \infty, n_1 \to \infty, p \to \infty, n \to \infty} \frac{1}{n_0} \to 0 < \frac{1}{n_1} \to 0 < \frac{1}{n_0} \to 0 < 0, \quad \lim_{p \to \infty}
\]

In this setting, a large body of work has been devoted to characterizing the expected true error and estimated error of LDA and its variants. The readers are referred to the paper by Raudys and Young who conducted a review of articles in this context.\(^{65}\) It is commonly assumed that the Mahalanobis distance, \( \delta_{\mu,p} = \sqrt{\left( \mathbf{\mu}_p - \mathbf{\mu}_p \right) \mathbf{\Sigma}_p^{-1} \left( \mathbf{\mu}_p - \mathbf{\mu}_p \right)^T} \), is finite and

\[
\lim_{p \to \infty} \delta_{\mu,p} = \delta_{\mu} < \infty
\]

where \( \delta_{\mu} \) denotes an arbitrary finite limiting point (see p. 4 of Refs. 67, 68, and 113). This condition ensures the existence of limits of performance metrics of relevant statistics.\(^{67,68}\) The aforementioned asymptotics along with the limiting point conditions is also referred to as “Kolmogorov asymptotics”.\(^{68}\) Raudys’s approximation of expected error rate of EDC that was presented in Ref. 25 is essentially based on the Kolmogorov asymptotics. As we discussed earlier, the result of the approximation of this type is far more accurate than its counterparts obtained from classical asymptotic settings (see Ref. 69 for a comparison). In the last two decades, the ideas of double asymptotics and, in particular, RMT have found eminent applications in engineering disciplines such as wireless communication. One of the first applications of this type of analysis in wireless communication was to characterize the performance of large multiuser linear receivers\(^{114}\) and to study various properties of code division multiple access (CDMA) channels where the number of users and the length of spreading code is large.\(^{115,116}\) These works paved the way for subsequent applications of RMT to study various properties of different communication channels; for instance, see Refs. 88, 117–120 to just cite a few articles.

\( G \)-analysis. According the general statistical analysis of observation (\( G \)-analysis) and its connection to \( G \)-estimation and the Kolmogorov asymptotic conditions, V. L. Girko, one of the pioneers in developing this theory, writes:

\[\text{The general statistical analysis of observations (}G\text{-analysis) is a mathematical theory studying some complex system } S, \text{ such that the number } m_n \text{ of parameters of its mathematical models can increase together with the growth of the number } n \text{ of observations of the system } S. \text{ The use of this theory consists in finding, with the help of observations of the system } S, \text{ mathematical models (}G\text{-estimators) that approach the system } S \text{ in some sense with a given rate under general assumptions on the observations: the existence of the distribution densities of the observed random vectors and matrices are not needed. The existence of several first moments of their components is all that is required; in addition, the numbers } m_n \text{ and } n \text{ satisfy the } G\text{-condition:}
\]

\[\lim_{n \to \infty} f \left( m_n, n \right) < \infty \quad (29)\]

where \( f(x,y) \) is some positive function increasing along \( x \) and decreasing along \( y \). In most cases the function \( f(m_n, n) \)
is equal to $xy^{-1}$. In this case the $G$-condition is also called the Kolmogorov condition.

The notation $\lim_{n \to \infty} f\left( m_{n}, n \right)$ in the comment above denotes $\lim \sup_{n \to \infty} f\left( m_{n}, n \right)$. By taking $f\left( m_{n}, n \right) = p/n$, the framework boils down to the “usual” double-asymptotic setting in which $p \to \infty$, $n \to \infty$, $p/n \to c$. Using this machinery and throughout a decade, Girko developed about 50 estimators for various purposes. They are named as $G_{1}$-estimator (an estimator of generalized variance), $G_{2}$-estimator (an estimator of Stieltjes transform of the normalized spectral function), $G_{3}$-estimator (estimator of inverse covariance matrix), etc. Among them, we observe an estimator of the solution of the discrete Kolmogorov–Wiener filter ($G_{b}$), and an estimator of Anderson statistics ($G_{13}$). A summary of his results is presented in.\(^{2}\) As pointed out by Girko, an important property of $G$-analysis is that the result does not depend on the actual distribution of data. An example was given in the previous section in which the convergence of empirical spectral distribution to M–P law is independent of distribution of data (Fig. 2). This provides us with a mathematical machinery not only to calibrate many traditional estimators from the standpoint of double asymptotics (the so-called generalized consistency\(^{3}\)) but also to synthesize distribution-free techniques.

In recent years, several research groups, predominantly in signal processing community, have utilized the idea of $G$-analysis in various settings. For example, Mestre and Lagunas derived the generalized consistent estimator of the optimum loading factor in spatial filtering.\(^{2}\) In Ref. 3, Rubio and Mestre used the $G$-analysis to first evaluate the performance of a global minimum variance portfolio (GMVP) implementation based on shrinkage covariance matrix estimation and weighted sampling. Then they used the $G$-analysis to characterize the limiting expression of the randomized variance and, based on that, they achieved a generalized consistent estimator of out-of-sample portfolio variance.\(^{1}\) In Ref. 121, the authors developed an estimator of the optimal linear filter for both multiantenna array signals and financial asset returns. In Ref. 5, we utilized the $G$-analysis to calibrate a traditional estimator of the true error of the regularized LDA. This classical estimator, known as the plug-in estimator, is consistent under $n \to \infty$ and fixed $p$ regime with a poor performance in small-sample situations. We observe that the calibrated new estimator can outperform not only the plug-in estimator but also other estimators of the true error, including Bootstrap 0.632 and cross-validation, in many situations in terms of bias and root-mean-square (RMS) error. Some other applications of $G$-analysis include estimating the eigenvalues and eigenvectors of sample covariance matrix\(^{122}\) and estimating the direction of arrival (DoA) in linear sensor arrays.\(^{123}\)

Extending $G$-analysis to Bayesian settings. In Ref. 124, we characterized the moments of a Bayesian minimum mean-square error (MMSE) error estimator, $\hat{\varepsilon}_{c}^{B}$,\(^{123,126}\) of the true error rate of LDA under a Gaussian model. There, we have characterized two sets of performance metrics: 1) the first, second, and cross moments of the estimated and actual errors conditioned on a fixed feature-label distribution, and 2) the unconditional moments. This means that we have obtained the metrics of performance of $\hat{\varepsilon}_{c}^{B}$ depending on the evaluation scheme—conditional moments (unconditional moments) can be used as if the estimator is evaluated in a frequentist (Bayesian) framework. We set up a series of conditions to which we refer as the Bayesian–Kolmogorov asymptotic conditions. These conditions allow us to characterize the performance metrics of Bayesian MMSE error estimation in an asymptotic sense. The Bayesian–Kolmogorov asymptotic conditions are set up based on the assumption of increasing $n$, $p$, and the certainty parameter $v$ with an arbitrary constant limiting ratio between $n$ and $p$, and $n$ and $v$. To our knowledge, these conditions permit, for the first time, the application of $G$-analysis in a Bayesian setting.

To analyze the Bayesian MMSE error estimator $\hat{\varepsilon}_{c}^{B}$, we consider a sequence of Gaussian discrimination problems as

$$\left( \mu_{p,0}, \mu_{p,1}, \Sigma_{p}, n_{p,0}, m_{p,0}, m_{p,1}, v_{p,0}, v_{p,1} \right), p = 1, 2, \ldots, \tag{30}$$

where $\Sigma$ and $\mu$ are the covariance matrix (assumed to be known) and the mean of Gaussian feature label distributions, respectively, and $\Sigma v$ and $m$ are the covariance matrix and the mean of the Gaussian prior distribution on $\mu$. We assume that the following limits exist for $ij = 0, 1$: \(\lim_{p \to \infty} m_{p,0}^{-1} \Sigma_{p,0,0}^{-1} \mu_{p,0} = m_{0}^{-1} \Sigma_{0}^{-1} \mu_{0,0}\), \(\lim_{p \to \infty} m_{p,0}^{-1} \Sigma_{p,0,0}^{-1} \mu_{p,0} = m_{0}^{-1} \Sigma_{0}^{-1} \mu_{0,0}\), and \(\lim_{p \to \infty} m_{p,0}^{-1} \Sigma_{p,1,0}^{-1} \mu_{p,1} = m_{0}^{-1} \Sigma_{0}^{-1} \mu_{0,1}\), \(\lim_{p \to \infty} m_{p,1}^{-1} \Sigma_{p,0,1}^{-1} \mu_{p,0} = m_{0}^{-1} \Sigma_{0}^{-1} \mu_{0,1}\), \(\lim_{p \to \infty} m_{p,1}^{-1} \Sigma_{p,1,1}^{-1} \mu_{p,1} = m_{0}^{-1} \Sigma_{0}^{-1} \mu_{0,1}\) and $\Sigma_{0,0}$ are some constants to which the limits converge. These are generalizations of $\lim_{p \to \infty} \delta_{p,0} = \delta_{0,0}$ condition that was previously stated in the Kolmogorov asymptotics context. All of the aforementioned conditions along with $v_{p} \to \infty$, $\gamma_{p} / n_{p} \to \gamma$, $\gamma \in (0, \infty)$, construct the Bayesian–Kolmogorov asymptotic conditions. To summarize, the asymptotic regime that we considered there is characterized by the following limit:

$$\lim_{p \to \infty, n_{p} \to \infty, \rho_{p} \to \infty} \frac{n_{p}}{\rho_{p}} m_{p,0}^{-1} \Sigma_{p,0,0}^{-1} \mu_{p,0} \to 0,1; \quad m_{p,0}^{-1} \Sigma_{p,0,0}^{-1} \mu_{p,0} \to m_{0}^{-1} \Sigma_{0}^{-1} \mu_{0,0}\)\(^{123}\)

$$m_{p,0}^{-1} \Sigma_{p,0,0}^{-1} \mu_{p,0} \to m_{0}^{-1} \Sigma_{0}^{-1} \mu_{0,0}$$

$$\delta_{p,0} / \rho_{p} \to 0,1; \quad \delta_{p,0} / \rho_{p} \to \delta_{0,0}\)\(^{123}\)

$$\delta_{p,0} / \rho_{p} \to 0,1; \quad \delta_{p,0} / \rho_{p} \to \delta_{0,0}\)\(^{123}\)

This limit is defined for a situation in which there is a conditioning on a specific value of feature label distribution parameters such as $\mu_{p,0}$. Therefore, in this case $\mu_{p,0}$ is not a random variable, and for each $p$, it is a vector of constants. Absent such conditioning, the sequence of discrimination problems and the above limit reduce to

$$\left( \Sigma_{p}, n_{p,0}, n_{p,1}, m_{p,0}, m_{p,1}, v_{p,0}, v_{p,1} \right), p = 1, 2, \ldots, \tag{32}$$
and

$$
\lim_{p \to \infty, n \to \infty, y_0 \to \infty} \frac{T}{n} \to j_0, 
\frac{\gamma}{\gamma} \to 0, 
\frac{\gamma}{\gamma} \to 1
$$

respectively. The accuracy of the closed-form expressions derived in this work is remarkable. We believe this framework can serve as a potential technique in the future to calibrate various Bayesian estimators from a frequentist point of view to wit, to optimize the performance of a developed Bayesian method for a specific parametric distribution with unknown parameters.

**Discussion**

In recent years, various statistical learning rules have been put forward for cancer diagnosis, prognosis, discriminating stages of cancer, types of pathology, and duration of survivability based on molecular profiles such as gene or protein expression patterns and single nucleotide polymorphism genotypes. Such a biomarker discovery process in high-throughput genomic and proteomic profiles has presented the statistical learning community with a challenging problem, namely how to learn from a large number of variables and a relatively small sample size. The properties of high-dimensional data, though, are not well understood. A high-dimensional setting is not the place to rely on intuition, nonrigorous propositions, and heuristics. At the same time, the classical notion of statistical consistency, which guarantees the performance of many classical statistical techniques, falters because this notion guarantees the performance of a technique in situations where the number of measurements unboundedly increases (n → ∞) for a fixed dimensionality of observations, p. In a finite sample operating regime, this implies that in order to expect an acceptable performance from a statistical technique, we need to have many more sample points than variables—a scenario opposite to what we currently face in high-throughput biology. Despite many achievements in the last few decades in the field of statistical learning, some of the most elementary problems remain unsolved. In this regard, Serdobolskii stated:

> It is difficult to describe the recent state of affairs in applied multivariate methods as satisfactory. Unimprovable (dominating) statistical procedures are still unknown except for a few specific cases. The simplest problem of estimating the mean vector with minimum quadratic risk is unsolved, even for normal distributions. Commonly used standard linear multivariate procedures based on the inversion of sample covariance matrices can lead to unstable results or provide no solution in dependence of data. Thus nearly all conventional linear methods of multivariate statistics prove to be unreliable or even not applicable to high-dimensional data.

Two mathematical–statistical machineries, discussed herein, show promising results in constructing techniques of high-dimensional data analysis: 1) shrinkage, and 2) Girko G-analysis. This paper presented a brief history of development, the underlying assumptions, and some of important results of each machinery. While in the last decade there has been some effort to create statistical software packages from some of the shrinkage methods, the methods developed through G-analysis remain mostly in the literature and unknown to many theoreticians and practitioners. Some effort from applied statistics and the signal processing community seems worthwhile in order to create ready-to-use software packages from these methods. In addition, practical implications of the underlying assumptions in G-analysis need further investigation. For example, we assumed n → ∞, p → ∞, p/n → c, 0 < c < ∞ along with some conditions on the existence of moments of random variables involved. In an asymptotic sense, the results are applicable to any ratio of p/n. However, in a finite-sample regime, it would be interesting to study the robustness of designed methods with respect to this ratio. Other natural directions that deserve further research in this line of work are 1) understanding the true nature of the so-called curse of dimensionality phenomenon, 2) the connection between G-analysis conditions and the curse of dimensionality, 3) the possibility of using G-analysis in creating lasso-like operators that have the capability of performing model selection, and 4) extending G-analysis to Bayesian statistics.

**Author Contributions**

Conceived the concepts: AZ. Wrote the first draft of the manuscript: AZ. Developed the structure and arguments for the paper: AZ. Made critical revisions: AZ. The author reviewed and approved of the final manuscript.

**REFERENCES**

1. Efron B. Bayesians, frequentists, and scientists. *J Am Stat Assoc*. 2005;100:15.
2. Mestre X, Lagunas MA. Finite sample size effect on minimum variance beamformers: optimum diagonal loading factor for large arrays. *IEEE Trans Signal Process*. 2006;54:69–82.
3. Rubio F, Mestre X, Palomar DP. Performance analysis and optimal selection of large minimum variance portfolios under estimation risk. *IEEE J Sel Topics Signal Process*. 2012;6:337–50.
4. Girko VL. *Statistical Analysis of Observations of Increasing Dimension*. Dordrecht: Kluwer Academic Publishers; 1995.
5. Zollanvari A, Dougherty ER. Generalized consistent error estimator of linear discriminant analysis. *IEEE Trans Signal Process*. 2015;63:2804–14.
6. Chandrasekaran B, Jain AK. Quantization complexity and independent measurements. *IEEE Trans Comput*. 1974;23:102–6.
7. Girko VL. *An Introduction to Statistical Analysis of Random Arrays*. Utrecht: VSP; 1998.
8. Pourahmadi M. *High-Dimensional Covariance Estimation*. New York City, NY: Wiley; 2013.
9. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*. 2005;365:488–92.
10. Hanczar B, Hu A, Sima C, Weinstein J, Bittner M, Dougherty ER. Small-sample precision of ROC-related estimates. *Bioinformatics*. 2010;26:822–30.
11. Michiels S, Koscielny S, Hill C. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst*. 2007;99:147157.
12. Robinson EB, Howrigan D, Yang J, et al. Response to predicting the diagnosis of autism spectrum disorder using gene pathway analysis. *J Autism Dev Disord*. 2014;44:1542–5.
13. Yousefi MR, Dougerty ER. Performance reproducibility index for classification. *Bioinformatics*. 2012;28:2824–33.
14. Dougerty ER, Braga-Neto U. Epistemology of computational biology: mathematical models and experimental prediction as the basis of their validity. *J Biol Syst*. 2006;14:65–90.
15. Dougerty ER. On the epistemological crisis in genomics. *Curr Genomics*. 2008;9:69–79.
16. Dougerty ER, Hua J, Bittner ML. Validation of computational methods in genomics. *Curr Genomics*. 2007;8:1–19.
17. Braga-Neto UM. Fads and fallacies in the name of small-sample microarray classification. *IEEE Signal Process Mag*. 2007;24:91–9.
18. Baranchik AJ. Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution. *J Am Stat Assoc*. 1977;68:117–30.
19. James W, Stein C. Estimation with Quadratic Loss. *Ann Math Statist*. 1962;3:1191.
20. James W, Stein C. Estimation with Quadratic Loss. *Ann Math Statist*. 1962;3:140–75.
21. Khotanzad A. A new approach to improving the James-Stein estimator. *J Multivar Anal*. 1994;50:356–69.
22. Brandwein AC, Straderman WE. Stein estimation: the spherically symmetric case. *Stat Sci*. 1990;5:363–69.
23. Brandwein AC, Straderman WE. Stein estimation for spherically symmetric distributions: recent developments. *Stat Sci*. 2012;27:1–13.
24. Gupta AK, Pena EA. A simple motivation for James-Stein estimator. *Stat Probab Lett*. 1991;12:337–40.
25. Hoffmann K. Stein estimation – a review. *Statistical Pap*. 2000;41:127–58.
26. Grother M. Improving Efficiency by Shrinkage: The James-Stein and Ridge Estimation Regressors. *New York, NY: CRC Press*. 1998.
27. Baranchik AJ. Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution. Stanford Univ, Stanford, Technical Report; 1964.
28. Baranchik AJ. A family of minimax estimators of the mean of a multivariate normal distribution. *Ann Math Stat*. 1970;41:642–5.
29. Hartle HL. Admissible estimators, recurrent distributions, and insoluble boundary problems. *Ann Math Stat*. 1971;42:855–903.
30. DasGupta A, Sinha BK. A New General Interpretation of the Stein Estimate and How it Adapts: with Applications. Purdue Univ, West Lafayette, Technical Report; 1997.
31. Guo YY, Pal N. A sequence of improvements over the James-Stein estimator. *J Multivar Anal*. 1991;36:121–62.
32. Kubokawa T. An approach to improving the James-Stein estimator. *J Multivar Anal*. 1994;50:356–69.
33. Hoerl AE. Application of Ridge analysis to regression problems. *Chem Eng Prog*. 1962;58:54–9.
34. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 1970;12:55–9.
35. Hoerl AE, Kennard RW. Ridge regression: applications to nonorthogonal problems. *Technometrics*. 1970;12:69–82.
36. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33:1–22.
37. Troppe JA. Just relax: convex programming methods for finding sparse signals in noise. *IEEE Trans Inf Theory*. 2006;20:1030–50.
38. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B*. 1996;58:267–88.
39. Donoho D, Johnstone I. Ideal spatial adaptation by wavelet shrinkage. *Biometr. J*. 1994;81:425–55.
40. Osborne M R, Presnell B, Turlach BA. On the LASSO and its dual. *J Comput Graph Stat*. 1999;9:3–37.
41. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat*. 2004;32:407–99.
42. Frank EK, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics*. 1993;35:109–35.
43. Efron B, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B*. 2005;67:301–20.
44. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101:1418–29.
45. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc B*. 2007;68:49–67.
46. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2007;101:235–50.
47. Mehta ML. On the statistical properties of the level-spacings in nuclear spectra. *Nucl Phys*. 1960;18:395–419.
85. Mehta ML, Gaudin M. On the density of eigenvalues of a random matrix. Nuel. Phys. 1960;18:420–7.
86. Mehta ML. Random matrices in nuclear physics and number theory. Contemp. Math. 1986;50:295–309.
87. Mehta ML. Determinants of quaternion matrices. J Math Phys Sci. 1974;8:592–9.
88. Hashem W, Khorunzhyy O, Loubaton P, Najim J, Pastur L. A new approach for capacity analysis of large dimensional multi-antenna channels. IEEE Trans Inf Theory. 2008;54:3987–4004.
89. Pastur L. A simple approach to global regime of the random matrix theory. Mathematical Results in Statistical Mechanics. 1999;429:54.
90. Girko VL. Circle law. Theory Probab Appl. 1984;4:694–706.
91. Girko VL. On the circle law. Theory Probab Math Statist. 1984;28:15–23.
92. Girko VL. Theory of Random Determinants. Dordrecht: Kluwe Academic Publishers; 1990.
93. Silverstein JW. Comments on a result of Yin, Bai and Krishnaiah for large dimensional multivariate F matrices. J Multivar Anal. 1984;15:408–9.
94. Silverstein JW, Bai ZD. On the empirical distribution of eigenvalues of large dimensional random matrices. J Multivar Anal. 1984;28:15–23.
95. Silverstein JW, Tuilino AM. Theory of large dimensional random matrices for engineers. In: IEEE Ninth International Symposium on Spread Spectrum Techniques and Applications, Manaus-Amazon; 2006:458–64.
96. Bai ZD, Yin YQ. On limiting spectral distribution of product of two random matrices when the underlying distribution is isotropic. J Multivar Anal. 1986;19:189–200.
97. Bai ZD, Yin YQ. On the convergence of the spectral empirical process of Wigner matrices. Bernoulli. 2005;11:1039–92.
98. Bai ZD, Saradassah E. Effect of high dimension comparison of significance tests for a high dimensional two sample problem. Stat Sin. 1996;6:311–29.
99. Bai ZD. Convergence rate of expected spectral distributions of large random matrices. Part I. Wigner matrices. Ann Probab. 1993;21:625–48.
100. Yin YQ, Bai ZD, Krishnaiah PR. Limiting behavior of the eigenvalues of a multivariate F matrix. J Multivariate Anal. 1983;13:508–16.
101. Yin YQ, Bai ZD, Krishnaiah PR. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. Probab Theory Relat Fields. 1988;78:509–21.
102. Yin YQ, Bai ZD, Krishnaiah PR. A limit theorem for the eigenvalues of product of two random matrices. J Multivar Anal. 1983;13:489–507.
103. Bai ZD, Yin YQ. Necessary and sufficient conditions for the almost sure convergence of the largest eigenvalue of Wigner matrices. Ann Probab. 1988;16:1729–41.
104. Raudys S. On the amount of a priori information in designing the classification algorithm. Tech Cybern. 1972;4:168–174. [In Russian].
105. Deev AD. Asymptotic expansions for distributions of statistics W, M, and W* in discriminant analysis. Stat Methods Classif. 1972;31:6–57. [In Russian].
106. Meshalkin LD, Serdoboltski VI. Errors in the classification of multi-variate observations. Theory Prob Appl. 1978;23:741–50.
107. Serdoboltski VI. On minimum error probability in discriminant analysis. Soviet Math Dokl. 1983;27:720–5.
108. Raudys S. Comparison of the estimates of the probability of misclassification. In: Proc. International Joint Conference on Pattern Recognit, Kyoto, Japan; 1978:280–2.