Title
Non-Gaussian Component Analysis

Author
Bean, Derek

Publication Date
2014

Non-Gaussian Component Analysis

by

Derek Merrill Bean

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter J. Bickel, Co-chair
Professor Noureddine El Karoui, Co-chair
Professor Laurent El Ghaoui

Spring 2014

Non-Gaussian Component Analysis

Abstract

Non-Gaussian Component Analysis

by

Derek Merrill Bean

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Peter J. Bickel, Co-chair

Professor Noureddine El Karoui, Co-chair

Extracting relevant low-dimensional information from high-dimensional data is a common pre-processing task with an extensive history in Statistics. Dimensionality reduction can facilitate data visualization and other exploratory techniques, in an estimation setting can reduce the number of parameters to be estimated, or in hypothesis testing can reduce the number of comparisons being made. In general, dimension reduction, done in a suitable manner, can alleviate or even bypass the poor statistical outcomes associated with the so-called "curse of dimensionality."

Statistical models may be specified to guide the search for relevant low-dimensional information or "signal" while eliminating extraneous high-dimensional "noise." A plausible choice is to assume the data are a mixture of two sources: a low-dimensional signal which has a non-Gaussian distribution, and independent high-dimensional Gaussian noise. This is the Non-Gaussian Components Analysis (NGCA) model. The goal of an NGCA method, accordingly, is to project the data onto a space which contains the signal but not the noise.

We conduct a comprehensive review of NGCA. We analyze the probabilistic features of the NGCA model and elucidate connections to similar well-known models and methods in the literature, including a hitherto-unseen and surprising connection to a set of models proposed by Cook in the context of dimension-reduction in regression. We review the literature on NGCA, catalogue existing NGCA methods, and compare them to the method proposed in Chapter 2.

We also propose and analyze a new NGCA method based on characteristic functions called CHFNGCA. We show CHFNGCA is, under mild moment conditions on the non-Gaussian sources, consistent and asymptotically normal; the latter property has not been demonstrated for any other NGCA method in the literature. We conclude by highlighting areas for future work.

The proof of all stated propositions, lemmas and theorems are contained in Appendices A and B.

To my family: my mother Lesa, my father Merrill, and my late sister Kasi.

For all your love and support.

Contents

Acknowledgments

Where to begin? First I would like to thank my advisors Dr. Peter Bickel and Dr. Noureddine El Karoui. Their guidance during this process was absolutely indispensable. Their insights substantially improved the final product; their time overseeing my research was beyond generous; and their patience while I've struggled is appreciated beyond what words can convey. The intellectual rigor they bring to bear on problems both challenges and inspires me to be better. I want to thank Peter for countless stimulating exchanges and for being a sympathetic ear during some difficult times in my life. And I want to thank Noureddine for the hours of high quality advice on Statistics, work and life. I can't possibly repay them for all they have given me.

I also want to thank Laurent El Ghaoui of my dissertation committee for his time and efforts. It is greatly appreciated.

I want to thank my outstanding research collaborators, Dr. Elmar Diederichs and Dr. Nima Reyhani, for many illuminating exchanges concerning NGCA. I also want to thank Dr. Chinghway Lim and Dr. Bin Yu for other collaborations not contained herein. Their contributions, in the form of stimulating discussions of ideas and thorough, careful work greatly improved those other endeavors,. I am deeply grateful for that.

The Berkeley Statistics Department is superb in all respects. It is a challenging, supportive, and rewarding place to be. It has truly been a pleasure to work here, and that is entirely due to the efforts of every member of the department. I want to thank the faculty. In particular, I would like to thank Dr. Ani Adhikari for helping me transition to living in Berkeley. Working as her teaching assistant for two semesters, and observing the high standards she set, made me a more effective teacher. I want to thank Dr. Martin Wainwright for rendering a great deal of very difficult material intelligible and fun to learn. I want to thank Dr. David Brillinger for many interesting discussions, particularly about ice hockey; and for being a great instructor to assist. I would like to thank the other faculty for whom I've been a teaching assistant: Dr. Nayantara Bhatnagar, Dr. Ching Shui-Cheng, Dr. Philip Stark, Mike Leong, and Dr. Adityanand Guntuboyina. Educating undergraduates with these fine instructors has been a rich and rewarding experience for me. I want to thank Dr. Jim Pitman for giving me a project along with some good advice my first summer here – I was in need of both. Finally, I want to thank Dr. Deb Nolan for a memorable week in New York City teaching modern statistical computing to smart, promising undergraduates. I am very lucky to have been a part of that.

My peers, past and present, in the Statistics graduate program have enhanced my work–and my life–immeasurably. They are creative, brilliant, talented and kind. They made the environment here exciting and supportive, and I have benefitted tremendously from their help in the courses, meetings and seminars we attended together. It is an honor to count them among my friends and acquaintances. I would like to recognize by name, in no particular order, Dr. Mike Higgins, Dr. Karl Rohe, Dr. Chris Haulk, Dr. Chinghway Lim, Dr. Brad Luen, Dr. Richard Liang, Dr. Dave Shilane, Dr. Megan Goldman, Dr. Jing Lei, Dr. Choongsoon Bae, Dr. Garvesh Raskutti, Dr. Nancy Wang, Laura Derksen, Dr.

Johann Gagnon-Bartsch, Dr. Yuval Benjamini (thanks for all the vigorous arguments!), Dr. James Long, Dr. Sean Ruddy, Dr. Wayne Lee, Dr. Winston Lin, Dr. Joe Neeman, Dr. Ngoc Tran, Dr. Oleg Mayba, Tamara Broderick, Luis Campos, Dr. Yu-Jay Huoh, Stephanie Sapp, Sharmodeep Bhattacharyya, Jeffrey Regier, and Steven Troxler. If I left anyone off this list I am deeply sorry: I still regard their acquaintance as a valuable gift.

I want to thank all the Department staff, without whose patience and helpfulness this would not have been possible. By name I would like to thank Angie Fong and La Shana Porlaris for repeatedly answering questions I should have probably already known the answers to. I want to thank Ryan Lovett for always taking the time to help me with computing issues, and for lending me a computer at the last minute on multiple occasions. Many thanks to Denise Yee for her helpfulness, kindness, and for letting me into my office those times I managed to lock myself out. And thanks to Mary Melinn – no matter what she was doing, she'd always take the time to help me when I needed something. Seemingly small considerate acts become enormous gestures for a student who started out so unsure of himself.

There are many people outside the department who were instrumental in my getting to this point. First and foremost I would like to thank Dr. Sundar Subramanian for giving me my first shot at statistical research as an undergraduate: there is no way I could have made it to Berkeley without his mentorship. I would like to thank all the other members of the Mathematics Department at the University of Maine I worked with for their excellent preparation. Thanks to all the hard-working teachers of the Messalonskee school district back in Oakland, Maine. Of these I'd like to single out Frank Brown in particular: thanks for teaching me how to think, and encouraging my lasting passion for ideas.

I could not have done it without my amazing friends. Thanks, Sara Taylor Rohe, for jokes and encouragement. Thanks to Kenda Templeton for her steadfast friendship these last two years. Thanks Samantha Shandorf for all the support. I especially want to acknowledge my closest friends in the world: Ryan Hross and his lovely wife Sarah, Jacob Burwood and his lovely wife Mary Beth, Mel Grout, Adam Stallilonis, Tyler Jolicoeur and Doug Putnam. It is a source of immense pride and satisfaction that I can count them among my friends. I cannot thank them enough for how much they have enriched my life through their humor, their restless and creative energy, and their kindness.

Last, but certainly not least, I want to acknowledge my wonderful family. This includes my parents Lesa and Merrill Bean especially. Their love and support; their encouragement for me to improve as a person; the comfort they generously gave to me with no hesitation during all those times I felt low: these things are the sine qua non of who I am today, and all I have achieved. I would like to acknowledge my mother's parents, Eileen Stevens and her late husband Leslie. I would also like acknowledge my father's late parents Vivian and Bernard Bean, whose lives touched so many hearts for the better. And finally, I want to acknowledge my late sister Kasi Gajtkowski and her husband Cliff Gajtkowski. Cliff was a good husband and is a good man, and supported my sister to the very end. Her recent passing is a deep source of sorrow and loss to all of us in the family; yet I am so lucky to have known her for the time I did. She was warm, creative, kind and funny. She is greatly missed, but she lights up my memories every day.

Chapter 1

Introduction to NGCA

1.1 Motivation

Sweeping technological advances in the acquisition and storage of data have produced datasets of ever-increasing size and complexity. The sheer quantity of data holds great promise: properly analyzed, these high-dimensional data, or "Big Data" as they are popularly known, may help researchers and data practitioners solve critical problems in the sciences, medicine, and finance. On the other hand, high-dimensional data are challenging to analyze. Only low-dimensional projections of the data can be visualized, and such projections may miss the important features. The statistician's broad aim of teasing out the stable structural component of the data and eliminating noise is beset by the fact that data tend to be sparsely distributed in high-dimensional space. Moreover, said structure could be complex, and the complex models called for by high-dimensional data require estimation of many parameters by comparatively few data points. The results, potentially, are highly variable, unstable, untrustworthy estimates. In a hypothesis testing context, many comparisons must be made, increasing the likelihood of false positives and "seeing whatever you want to see" in the data. In sum, there is an apparent "Curse of Dimensionality."

One solution to this problem is to assume that the key statistical information lies on some low-dimensional structure. If this low-dimensional structure can be identified, data points can be projected onto it without loss of statistical information. Common statistical procedures can be performed on the low-dimensional data, bypassing the "curse."

Dimensionality reduction as a pre-processing step for analyzing high-dimensional data has a long history in Statistics. By far the oldest and most famous method for dimensionality-reduction is Principal Components Analysis (PCA) [36][25]. In PCA, the data are rotated such that the maximal amount of variance in the data is distilled along the new coordinate axes; typically only a small subset of the high-variance directions are retained. This amounts to a spectral decomposition of the covariance matrix, which, implicitly, is assumed to carry the important structure in the data. In the several decades following the introduction of PCA, many other methods for dimensionality reduction have been proposed to obtain "good" low-dimensional representations of the data.

Linear projections are desirable for their simplicity. A plausible model for high-dimensional data can guide the determination of projection directions which retain the "signal" in the data while eliminating the "noise." These observations are at the heart of Non-Gaussian Components Analysis (hereafter referred to as NGCA) [30][5][31][38][29][32][15][35][37][16]. Two ideas underlie NGCA: 1) it is realistic to model the data generating process as a mixture of independent sources, the idea behind Independent Component Analysis (ICA) [9][27]; 2) the important structure in the data is non-Gaussian, the idea behind Projection Pursuit [20][26]. The NGCA model can be specified by the NGCA decomposition:

Definition 1.1.1 (NGCA Decomposition.)**.** *We say a p-dimensional random vector X has a d-dimensional NGCA decomposition ($d < p$) if there exists a $p \times d$ matrix Γ and a $p \times (p-d)$ matrix η such that:*

$$\begin{bmatrix} \Gamma^T X \\ \eta^T X \end{bmatrix} = \begin{bmatrix} V \\ G \end{bmatrix} \tag{NGCA Decomposition}$$

where the random vector $V \in \mathbb{R}^d$ is non-Gaussian and independent of the $(p-d)$-dimensional Gaussian vector G.

In the NGCA setup, we assume the observations are i.i.d. copies of the data vector X. Estimation of Γ or η is not possible, because they are not identifiable: for any $d \times d$ full-rank matrix A, if we set $\Gamma_1 = \Gamma A$, then $(\Gamma_1)^T X = AV$ is non-Gaussian and independent of the Gaussian component (the same reasoning applies to η). However, the column spaces of Γ and Γ_1 coincide. Thus, the goal of NGCA is to estimate the subspace spanned by the columns of Γ. We call this subspace the *non-Gaussian subspace*. The subspace generated by the columns of η is called the *Gaussian subspace*.

This chapter is an overview of the properties of NGCA models and of past approaches for estimating the non-Gaussian subspace. In Section 2 we review some probabilistic features of the NGCA model. In Section 3 we examine a surprising connection between NGCA and a regression model proposed by Cook [10][12] in the context of sufficient dimension reduction in regression [11][1]. In Section 4 we provide a detailed review of past approaches to NGCA. Proofs are contained in Appendix A.

1.2 The NGCA model: assorted examples.

The goal of this section is to acquaint the reader with various features of the NGCA model. All of the results contained herein have appeared before in various places in the NGCA literature, but we have compiled the most important ones together. Familiarity with the model is a prerequisite for understanding subsequent sections of this dissertation and engaging with the NGCA literature at large. Therefore, a systematic investigation of properties of the NGCA model under various assumptions could prove useful. We adopt a pedagogical style for clarity.

Invertibility of $[\Gamma \quad \eta]$.

In the specification of the NGCA model via Definition 1.1.1 we do not explicitly state a structural stochastic model for X; we only have a model for X after it is transformed by the "recovery matrix" $[\Gamma \quad \eta]^T$ (so-named because it recovers the independent components). We did not specify that this matrix is invertible, in which case we could immediately write down the form of X. However, it can be shown that when $[\Gamma \quad \eta]^T$ is singular (non-invertible) the non-Gaussian component has certain undesirable properties. Thus, imposing an invertibility assumption is no more restrictive than ruling out said undesirable properties. We show that when $[\Gamma \quad \eta]^T$ is invertible, the NGCA decomposition from Definition 1.1.1 is equivalent to the "generative model" for NGCA introduced in [38] and [37]. Then we show how invertibility can be used to derive a representation for the distribution of X, and deduce a Stein-like identity for NGCA models under smoothness assumptions on the non-Gaussian density.

Singularity of the matrix $[\Gamma \quad \eta]^T$ entails a poorly-behaved NGCA model, as the following proposition shows:

Proposition 1.2.1. *Let X be a p-dimensional random vector with a NGCA decomposition as in Definition 1.1.1. Suppose the recovery matrix $[\Gamma \quad \eta]^T$ is singular (i.e. its inverse does not exist). Then one of the following must be true:*

1. $\dim(\mathrm{span}(\Gamma)) < d$;
2. $\dim(\mathrm{span}(\eta)) < p - d$;
3. $\dim(\mathrm{span}(\Gamma) \cap \mathrm{span}(\eta)) > 0$, *which implies there exists $c \in \mathbb{R}^d$ with $c \neq 0$ such that $c^T V = k$ for some constant k with probability 1.*

In light of Proposition 1.2.1, to ensure nonsingularity of the recovery matrix, it is enough to assume that Γ and η are full rank, and that $\mathrm{Var}(c^T V) > 0$ for all nonzero $c \in \mathbb{R}^d$ (setting the variance to ∞ whenever it does not exist). Although we can choose to directly enforce the condition that the column spaces of Γ and η do not intersect, the condition $\mathrm{Var}(c^T V) > 0$ for all nonzero $c \in \mathbb{R}^d$ has an intuitive appeal in terms of non-Gaussian decompositions: it precludes the possibility that V itself has a NGCA decomposition with Gaussian components of variance 0 (where we identify a point mass at k with the $\mathcal{N}(k, 0)$ distribution). We shall see later in Theorem 1.2.12 that, when the covariance of X exists, the identifiability of the non-Gaussian subspace is equivalent to V having no NGCA decomposition with independent Gaussian components of any variance, 0 or otherwise. Thus, invertibility is actually a weaker condition than identifiability; and we generally require identifiability at a minimum to estimate the non-Gaussian space. At any rate, Proposition 1.2.1 provides a justification for only considering invertible recovery matrices.

Thus, for all but a small class of pathological NGCA models, the matrix $[\Gamma \quad \eta]^T$ which recovers the independent non-Gaussian and Gaussian components is invertible. It is therefore reasonable to just assume invertibility. We formalize this in another definition.

Definition 1.2.2. *We say the p-dimensional random vector X has an invertible NGCA decomposition if it has an NGCA decomposition as in Definition 1.1.1 such that the recovery matrix $[\Gamma \quad \eta]$ is invertible.*

We introduce now another way of formulating NGCA models. We call them "generative" NGCA models:

Definition 1.2.3. *We say the p-dimensional random vector X has a generative NGCA model if there exists $\bar{\Gamma} \in \mathbb{R}^{p\times d}$ and $\bar{\eta} \in \mathbb{R}^{p\times(p-d)}$ such that the $p \times p$ matrix $\begin{bmatrix}\bar{\Gamma} & \bar{\eta}\end{bmatrix}$ is invertible, and X can be written as*

$$X = \bar{\Gamma}V' + \bar{\eta}G',$$

where $V' \in \mathbb{R}^d$ is non-Gaussian, $G' \in \mathbb{R}^{p-d}$ is Gaussian, and V' is independent of G'.

This formulation of the NGCA model was used in [38] and [37]. Generative NGCA models are equivalent to NGCA models specified in Definition 1.2.2:

Proposition 1.2.4. *A random vector X has an invertible NGCA decomposition (Definition 1.1.1) with invertible recovery matrix $[\Gamma \quad \eta]$ if and only if X has the form*

$$X = \bar{\Gamma}V' + \bar{\eta}G'$$

where V' is a d-dimensional non-Gaussian random vector independent of $(p-d)$-dimensional Gaussian vector G', and the $p \times p$ matrix $\begin{bmatrix}\bar{\Gamma} & \bar{\eta}\end{bmatrix}$ is invertible. Furthermore, $\mathrm{span}(\bar{\Gamma})^\perp = \mathrm{span}(\eta)$ *and* $\mathrm{span}(\bar{\eta})^\perp = \mathrm{span}(\Gamma)$.

The upshot of Proposition 1.2.4 is that specifying a NGCA model from Definition 1.2.2 or from Definition 1.2.3)are equivalent when the non-Gaussian and Gaussian spaces are "well-behaved." Ultimately, how one chooses to specify the model becomes a matter of taste. We tend to prefer as our definition the decomposition displayed in Definition 1.1.1 because it defines the non-Gaussian space directly and makes its role more transparent. Most of the literature on NGCA begins with a submodel of Definition 1.2.3 which we call the "Non-Gaussian signal in Gaussian noise" model. Much of Section 1.2 is devoted to reconciling these different points of view, and showing that results which are obtained under one point of view transfer easily to others.

Invertibility of $[\Gamma \quad \eta]$ allows us immediately to write down a representation of the probability distribution of NGCA models. Under mild assumptions an identity similar to Stein's identity follows from this representation. This identity is the basis of many methods for estimating the non-Gaussian space.

Example: The distribution of NGCA models and a Stein-like identity. Suppose X has an invertible NGCA decomposition (Definition 1.2.2). Let F be the distribution of the non-Gaussian component V. Then we can represent the distribution P of X by

$$dP(x) = \left(\det(\Gamma\Gamma^T + \eta\eta^T)\right)^{\frac{1}{2}} dF(\Gamma^T x) d\Phi_{\mu_G, \Delta_G},$$

where $\mu_G = \mathbb{E}(G)$, $\Delta_G = \mathrm{Cov}(G)$ and $\Phi_{\mu,\Delta}$ is the $\mathcal{N}(\mu, \Delta)$ distribution function.

If we assume that V has a density function $f(x)$ on $\mathbb{R}^d$ then $[\Gamma \quad \eta]$ must be invertible due to Proposition 1.2.4, since the the set $\{x | c^T x = k\}$ for each $c \neq 0$ and each k has measure 0 under the Lebesgue measure on $\mathbb{R}^d$. If we further assume $\Delta_G \succ 0$ (i.e. Δ_G is invertible) then X has density $p(x)$ given by :

$$p(x) = \left(\det(\Gamma\Gamma^T + \eta\eta^T)\right)^{\frac{1}{2}} f(\Gamma^T x)\phi_{\mu_G,\Delta_G}(\eta^T x), \tag{1.1}$$

where $\phi_{\mu,\Delta}$ is the density function of the $\mathcal{N}(\mu, \Delta)$ distribution. This leads to a Stein-like identity, stated in the following proposition.

Proposition 1.2.5. *Assume $X \sim p(x)$ where $p(x)$ is defined in (1.1) with $\mu_G = 0$; assume the non-Gaussian density $f(x)$ is differentiable. Let g be a differentiable function on $\mathbb{R}^p$. Then provided we can differentiate under the integral sign,*

$$\mathbb{E}\left[\nabla g(X)\right] - \eta\Delta_G^{-1}\eta^T\mathbb{E}\left[Xg(X)\right] \in \mathrm{span}(\Gamma).$$

Stein's identity states $\mathbb{E}\left[\nabla g(X)\right] - \Sigma^{-1}\mathbb{E}\left[Xg(X)\right] = 0$ if and only if X is distributed $\mathcal{N}(0, \Sigma)$ with $\Sigma \succ 0$. However, we purposely obtained our identity under weaker conditions: for $X \sim p(x)$, where $p(x)$ is given in (1.1), we have not made the assumption $\Sigma = \mathrm{Cov}(X)$ exists.

If $\mathbb{E}\left[Xg(X)\right] = 0$ then the vector $\mathbb{E}\left[\nabla g(X)\right]$ lies in the non-Gaussian space. For linear functions g, imposing this condition forces $g \equiv 0$, hence g contains no information about the non-Gaussian space. However, nonlinear choices of g yield nontrivial vectors. If we estimate $\mathbb{E}\left[\nabla g(X)\right]$ by its empirical counterpart based on n i.i.d. samples for several choices of g, we can collect a group of vectors which lie close to the non-Gaussian subspace (up to estimation errors). This is the basic idea underpinning many approaches to estimating the target space.

Non-Gaussian signal in Gaussian noise model.

The version of the NGCA model found most commonly in the literature (e.g. [30][5][31][32][29][15][16]) is the "non-Gaussian signal in Gaussian noise model":

Definition 1.2.6. *We say a p-dimensional random vector X follows the non-Gaussian signal in Gaussian noise model if it can be written in the form*

$$X = \bar{\Gamma}S + N,$$

where $\bar{\Gamma}$ is a $p \times d$ matrix, S is a non-Gaussian d-dimensional random vector (the "signal"), and N is a p-dimensional random variable, independent of V', with distribution $\mathcal{N}(0, \Delta)$ for $\Delta \succ 0$.

The fully-dimensional Gaussian noise in Definition 1.2.6 is probably a more physically realistic model than the noise in the generative NGCA model in Definition 1.2.3. However the non-Gaussian signal in Gaussian noise model is a submodel of the generative NGCA model:

Proposition 1.2.7. *If a p-dimensional random vector X is distributed according to Definition 1.2.6, then it can be written in the form*

$$X = \bar{\Gamma}S + N_1 + N_2,$$

where $N_1 \in \mathrm{span}(\bar{\Gamma})$, $N_2 \in \Delta\mathrm{span}(\bar{\Gamma})^{\perp}$, and N_1 and N_2 are independent.

Clearly we can write $\bar{\Gamma}S + N_1 = \bar{\Gamma}V'$ for a d-dimensional non-Gaussian vector V' whose distribution is a convolution of a non-Gaussian distribution with a Gaussian; also, we can write $N_2 = \bar{\eta}G'$ for $\eta \in \mathbb{R}^{p\times(p-d)}$ with $\mathrm{span}(\eta) = \Delta\mathrm{span}(\bar{\Gamma})^{\perp}$ and G' a $p-d$-dimensional Gaussian vector independent of V'. This shows the non-Gaussian signal in Gaussian noise model (Definition 1.2.6) is a submodel of the generative NGCA model (Definition 1.2.3). The non-Gaussian space in this model is $\left(\Delta\mathrm{span}(\bar{\Gamma})^{\perp}\right)^{\perp} = \Delta^{-1}\mathrm{span}(\bar{\Gamma})$. The goal of NGCA is to eliminate as much of the independent Gaussian noise, represented by $N_2 = \bar{\eta}G'$, as possible. The distribution of the non-Gaussian variable V' inherits smoothness properties from its convolution with the Gaussian: for instance, the density exists and is differentiable. Thus the Stein-like identity (Proposition 1.2.5) underpinning many NGCA methods holds automatically (though we will recast it in slightly different form as we proceed).

Example: $\Delta = \sigma^2 I_p$. If the noise covariance Δ is proportional to the $p \times p$ identity matrix I_p, the non-Gaussian and Gaussian spaces must be orthogonal. Indeed, $\mathrm{span}(\Gamma) = \Delta^{-1}\mathrm{span}(\bar{\Gamma}) = \mathrm{span}(\bar{\Gamma})$ and $\mathrm{span}(\eta) = \mathrm{span}(\bar{\Gamma})^{\perp} = \mathrm{span}(\bar{\Gamma})$. However, this simple case is generally uninteresting from a NGCA perspective. The covariance model for the noise is unrealistic. Furthermore, the model becomes equivalent to a classical Factor Analysis model, and therefore the non-Gaussian space can be recovered as the subspace spanned by the leading d principal component directions (provided the covariance of S exists). In NGCA, the standard way to ensure the non-Gaussian space is orthogonal to the Gaussian space is to transform X to "whiten" the space, making $\Sigma = \mathrm{Cov}(X) = I_p$.

Example: Non-Gaussian signal in Gaussian noise: density and a Stein identity. Because the distribution of the non-Gaussian component has a density, the non-Gaussian signal in Gaussian noise model has a density of the form (1.1). There is another form of the density commonly given in the NGCA literature (e.g. [5][15][16]):

Proposition 1.2.8. *The p-dimensional random vector X distributed according to the non-Gaussian signal in Gaussian noise model (Definition 1.2.6) has a probability density p of the form*

$$p(x) = h(\Gamma^T x)\phi_\Delta(x),$$

where h is a differentiable real-valued function on $\mathbb{R}^d$, Γ is a $p \times d$ matrix such that $\mathrm{span}(\Gamma) = \Delta^{-1}\mathrm{span}(\bar{\Gamma})$, and ϕ_Δ is the density of the $\mathcal{N}(0, \Delta)$ distribution.

There are many ways to prove this proposition (see e.g. [5], Appendix A.1). Our proof is based on the classical statistical notion of sufficiency, as captured by the following lemma:

Lemma 1.2.9. *Let X be distributed according to the non-Gaussian signal in Gaussian noise model (Definition 1.2.6). For any $p \times d$ matrix Γ which satisfies $\text{span}(\Gamma) = \Delta^{-1}\text{span}(\bar{\Gamma})$, the conditional distribution of $X|\left(\bar{\Gamma}S = s, \Gamma^T X = t\right)$ does not depend on the value of s.*

Lemma 1.2.9 has an interesting interpretation in terms of sufficient statistics. Conditional on $\bar{\Gamma}S = s$, X has a normal distribution with mean s. In the context of estimating the mean parameter of a normal distribution with known covariance Δ, where the mean is assumed to lie in a known linear subspace, it follows that $\Gamma^T X$ is a sufficient statistic for estimating s in the classical sense. Of course, in the NGCA context, $\bar{\Gamma}S$ is random and Δ is unknown: we use sufficiency as a device to get a representation for the density of X via the factorization criterion for sufficiency ([33], p. 35, Theorem 6.5).

With the form of the density of the non-Gaussian signal in Gaussian noise model given in Proposition 1.2.8 we immediately deduce another Stein-like identity for such models:

Proposition 1.2.10. *Let X be distributed according to the non-Gaussian signal in Gaussian noise model (Definition 1.2.6) with density $p(x)$. Let g be a differentiable function on $\mathbb{R}^p$. Then provided we can differentiate under the integral sign,*

$$\mathbb{E}\left[\nabla g(X)\right] - \Delta^{-1}\mathbb{E}\left[Xg(X)\right] \in \text{span}(\Gamma).$$

For the purposes of estimating vectors lying in the non-Gaussian subspace span(Γ) we might be tempted to loosen the assumption $\mathbb{E}\left[Xg(X)\right] = 0$ and instead use an estimate of Δ^{-1}. However, Δ is not an identifiable parameter. (To see this, recall $X = \bar{\Gamma}S + N$. But we can always write $N = N' + N''$ where N', N'' are independent Gaussian vectors with $\text{Cov}(N') = \text{Cov}(N'') = \Delta/2$. Apply Proposition 1.2.7 to N' only to produce $N' = N_1' + N_2'$, with $N_1' \in \text{span}(\bar{\Gamma})$. Then $N_2' + N''$ is a Gaussian with strict positive definite covariance different from Δ.) We can bypass this problem by assuming $\Sigma = \text{Cov}(X)$ exists (which is equivalent to assuming Cov(S), the covariance of the non-Gaussian vector, exists). Since $\Sigma = \bar{\Gamma}\text{Cov}(S)\bar{\Gamma}^T + \Delta \succeq \Delta$, we immediately conclude Σ^{-1} also exists. Using the first resolvant matrix identity $A^{-1} - B^{-1} = A^{-1}(A-B)B^{-1}$ for invertible matrices A and B, and plugging in $A = \Delta$ and $B = \Sigma$, we obtain:

$$\Delta^{-1} = \Sigma^{-1} + \Delta^{-1}\bar{\Gamma}\text{Cov}(S)\bar{\Gamma}^T\Sigma^{-1};$$

from which we can write:

$$\mathbb{E}\left[\nabla g(X)\right] - \Delta^{-1}\mathbb{E}\left[Xg(X)\right] = \mathbb{E}\left[\nabla g(X)\right] - \Sigma^{-1}\mathbb{E}\left[Xg(X)\right] - \Delta^{-1}\bar{\Gamma}\text{Cov}(S)\bar{\Gamma}^T\Sigma^{-1}\mathbb{E}\left[Xg(X)\right].$$

Since $\text{span}(\Gamma) = \Delta^{-1}\text{span}(\bar{\Gamma})$ we can use Proposition 1.2.10 and conclude

$$\mathbb{E}\left[\nabla g(X)\right] - \Sigma^{-1}\mathbb{E}\left[Xg(X)\right] \in \text{span}(\Gamma)$$

(the same reasoning also shows $\text{span}(\Gamma) = \Delta^{-1}\text{span}(\bar{\Gamma}) = \Sigma^{-1}\text{span}(\bar{\Gamma})$). Since Σ is estimable we could then drop the assumption $\mathbb{E}[Xg(X)] = 0$ and still obtain estimates of vectors which lie in the non-Gaussian space. This result holds for the more general NGCA models of Definition 1.1.1 if we assume Σ exists, Σ^{-1} exists, the Gaussian component G has mean 0, and the non-Gaussian component V has a differentiable density.

Existence of $\Sigma = \text{Cov}(X)$ and the identifiability condition.

We now explore the probabilistic features of the NGCA model of Definition 1.1.1 and its variants under the assumption that $\Sigma = \text{Cov}(X)$ exists. This is equivalent to assuming that the covariance of the non-Gaussian component V exists (we do not consider Gaussian components of infinite variance). We discuss the relationship between NGCA and Principal Component Analysis (PCA). We then provide a theorem which characterizes NGCA models which have identifiable non-Gaussian and Gaussian spaces; this theorem assumes the existence of the covariance.

Example: NGCA and PCA. In Principal Component Analysis (PCA) [36][25], subspaces in which the data have maximal variance are considered informative. Accordingly, an eigen-decomposition of the sample covariance matrix is performed, and the data are projected onto the subspace spanned by the eigenvectors with the largest eigenvalues. NGCA, however, seeks the subspace in which the data are non-Gaussian, which does not necessarily coincide with the most variable directions. In fact, PCA may pick the Gaussian directions. We illustrate this in the following example; to simplify matters, we examine PCA performed on the population covariance matrix Σ.

Suppose X is a p-dimensional random vector distributed according to the NGCA generative model (Definition 1.2.3). Let $\bar{\Gamma}$ be a $p \times d$ orthogonal matrix, i.e. $\bar{\Gamma}^T\bar{\Gamma} = I_d$. Similarly, let $\bar{\eta}$ be a $p \times (p-d)$ orthogonal matrix, whose columns are orthogonal to $\bar{\Gamma}$, i.e. $\bar{\Gamma}^T\bar{\eta} = 0$. Also, assume $\text{Cov}(V') = D_{V'}$ and $\text{Cov}(G') = D_{G'}$ are diagonal. Then,

$$\Sigma = \bar{\Gamma} D_{V'} \bar{\Gamma}^T + \bar{\eta} D_{G'} \bar{\eta}^T.$$

Note that $\Sigma\bar{\Gamma} = \bar{\Gamma}D_{V'}$ and $\Sigma\bar{\eta} = \bar{\eta}D_{G'}$, which implies the columns of $\bar{\Gamma}$ and $\bar{\eta}$ are the eigenvectors of Σ, with eigenvalues $\text{Var}(V_i')$, $\text{Var}(G_j')$, $i = 1, \ldots, d$, $j = 1, \ldots, p-d$. If $\text{Var}(V_i') \geq \text{Var}(G_j')$ for all i and j, then the leading principal component directions correspond to the non-Gaussian subspace. However, we could have $\text{Var}(V_i') \leq \text{Var}(G_j')$ for all i and j, making the NGCA space correspond to the principal component directions with the smallest eigenvalues! Furthermore, we can find an arrangement of the $\text{Var}(V_i')$ and $\text{Var}(G_j')$ such that the NGCA space corresponds to any size d subset of the principal component directions. And if $\bar{\Gamma}$ and $\bar{\eta}$ are not orthogonal, the non-Gaussian space may not lie in the span of any subset of d principal directions. In NGCA, we simply do not make any assumptions about the variability of the Gaussian component relative to the non-Gaussian component. The sole criterion for determining interesting directions is non-Gaussianity.

Example: identifiability of the non-Gaussian and Gaussian subspaces. For the statistical problem of estimating the non-Gaussian space using a sample from a NGCA model, we require at minimum that the non-Gaussian space be an identifiable parameter from that model. This makes estimation possible. It turns out that we have simple necessary and sufficient conditions for identifiability, assuming that Σ exists.

Suppose $X \sim P$ is a NGCA model as in Definition 1.1.1. We define identifiability:

Definition 1.2.11 (Identifiable NGCA model)**.** *We say P is an identifiable d-dimensional NGCA model if for all other $\Gamma_1 \in \mathbb{R}^{p\times d}$ and $\eta_1 \in \mathbb{R}^{p\times(p-d)}$ such that $\Gamma_1^T X$ is a non-Gaussian vector independent of Gaussian vector $\eta_1^T X$, we must have $\operatorname{span}(\Gamma_1) = \operatorname{span}(\Gamma)$ and $\operatorname{span}(\eta_1) = \operatorname{span}(\eta)$. That is, any other d-dimensional NGCA decomposition has the same non-Gaussian and Gaussian spaces.*

If the identifiability condition is violated an estimator may not capture the important structure in the data. Consider the toy model $X = (X_1, X_2, X_3)$, where X_1 is non-Gaussian, and independent of (X_2, X_3) which are independent Gaussians. Clearly, (X_1, X_2) is non-Gaussian and independent of the Gaussian X_3. The span of the first two coordinates is thus a non-Gaussian subspace, and the span of the third coordinate is an independent Gaussian subspace. However, (X_1, X_2) carries undesirable Gaussian noise. Moreover, (X_1, X_3) is non-Gaussian and independent of Gaussian X_2 – but this decomposition corresponds to different non-Gaussian and Gaussian subspaces! The most useful decomposition is X_1 and (X_2, X_3) - a decomposition in which the dimension of the Gaussian component is in some sense maximal. Intuitively, the Gaussian component being in some sense maximal is a requirement for identifiability. This intuition is confirmed by the following theorem, which gives necessary and sufficient conditions for identifiability:

Theorem 1.2.12. *Let $X \sim P$ be a p-dimensional random vector with a NGCA decomposition as in Definition 1.1.1. Assume $\Sigma = \operatorname{Cov}(X)$ exists. Then the following are equivalent:*

(i) The non-Gaussian component V does not itself have an invertible d'-dimensional NGCA decomposition as in Definition 1.2.2 with $0 \le d' < d$.

(ii) The random vector X is distributed as a generative NGCA model (Definition 1.2.3) of the form:

$$X = \bar{\Gamma} V' + \bar{\eta} G'.$$

Furthermore, there does not exist a full rank $d \times d$ matrix M such that the first coordinate of MV' has a marginal Gaussian distribution independent of the other $d-1$ coordinates.

(iii) X has a d-dimensional NGCA decomposition such that the non-Gaussian and Gaussian subspaces are identifiable. That is, another d-dimensional NGCA decomposition will have the same non-Gaussian and Gaussian subspaces.

The theorem is adapted from Theorem 1.3 in [37]. The assumption that Σ exists is not a necessary condition for identifiability: both (i) and (ii) follow readily from (iii) without such an assumption. It is a sufficient condition however: it is used to show (ii)$\Rightarrow$(iii) in [37]. There, the Hessian of the characteristic function is used to characterize distributions with independent Gaussian components; the existence of the second derivatives of the characteristic function is guaranteed by the existence of the covariance matrix.

Existence of Σ^{-1}.

Example: when does Σ^{-1} not exist? Suppose in the NGCA model of Definition 1.1.1 the covariance $\Sigma = \mathrm{Cov}(X)$ exists and the identifiability conditions of Theorem 1.2.12 hold. The existence of Σ implies the existence of $\mathrm{Cov}(V)$; computing the covariance of X we have the relationship:

$$\begin{bmatrix} \Gamma^T \\ \\ \eta^T \end{bmatrix} \Sigma \begin{bmatrix} \Gamma & \eta \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}(V) & 0 \\ \\ 0 & \mathrm{Cov}(G) \end{bmatrix}. \tag{1.2}$$

Identifiability implies there does not exist $c \in \mathbb{R}^d$ such that $\mathrm{Var}(c^T V) = 0$. Otherwise, $c^T V = k$ with probability 1 for some constant k, and V contains a trivial independent Gaussian component (where k is taken to have a $\mathcal{N}(k, 0)$ distribution). Two conclusions follow: (1) if Γ and η are full rank, then by Proposition 1.2.1 the matrix $[\Gamma \quad \eta]$ is invertible. This implies Σ is invertible if and only if $\mathrm{Cov}(V)$ is invertible and $\mathrm{Cov}(G)$ is invertible. (2) For all c, we have:

$$0 < \mathrm{Var}(c^T V) = c^T \mathrm{Cov}(V) c,$$

which implies $\mathrm{Cov}(V) \succ 0$. Therefore if the identifiability holds, Σ is singular if and only if $\mathrm{Cov}(G)$ is singular.

In the non-Gaussian signal in Gaussian noise model (Definition 1.2.6) we have $\Sigma \succeq \Delta \succ 0$; therefore, if Σ exists, Σ^{-1} exists.

Example: Σ and the relationship between the non-Gaussian and Gaussian subspaces. So far we have made no assumptions about the relationship between the non-Gaussian and Gaussian subspaces. However, the special structure of the NGCA decomposition forces them to be orthogonal complements of one another in the inner-product defined by Σ.

Recall equation (1.2). Compute the left hand side and obtain the relation:

$$\begin{bmatrix} \Gamma^T \Sigma \Gamma & \Gamma^T \Sigma \eta \\ \\ \eta^T \Sigma \Gamma & \eta^T \Sigma \eta \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}(V) & 0 \\ \\ 0 & \mathrm{Cov}(G) \end{bmatrix}. \tag{1.3}$$

We see that $\Gamma^T \Sigma \eta = 0$ for all Γ, η that span the non-Gaussian and Gaussian subspaces respectively. Clearly $\Sigma\,\mathrm{span}(\eta) \subseteq \mathrm{span}(\Gamma)^{\perp}$. If Σ is not invertible, the relation is one of strict subset. But if Σ is invertible, the relation holds with equality. We state this as a proposition.

Proposition 1.2.13. *Let $X \in \mathbb{R}^p$ have an invertible NGCA decomposition as in Definition 1.2.2. Suppose $\Sigma = \mathrm{Cov}(X)$ exists with $\Sigma \succ 0$. Then $\Sigma\,\mathrm{span}(\eta) = \mathrm{span}(\Gamma)^\perp$ and $\mathrm{span}(\Gamma) = \Sigma^{-1}\mathrm{span}(\eta)^\perp$.*

If $\Sigma \succ 0$, instead of parameterizing the NGCA model by two independent subspace parameters (Γ, η), we can parameterize it by the non-Gaussian space Γ and covariance Σ. This is an appealing parameterization since Σ is readily estimable.

In the generative NGCA model (Definition 1.2.3) recall that $\mathrm{span}(\eta)^\perp = \mathrm{span}(\bar{\Gamma})$ (see Proposition 1.2.4). Therefore, we have the relation $\mathrm{span}(\Gamma) = \Sigma^{-1}\mathrm{span}(\bar{\Gamma})$. For the non-Gaussian signal in Gaussian noise model, we previously derived the relation $\mathrm{span}(\Gamma) = \Delta^{-1}\mathrm{span}(\bar{\Gamma})$ without assuming the existence of Σ (see Proposition 1.2.7). If Σ does exist, we can replace Δ^{-1} with Σ^{-1} in the relation. Using Σ is preferred, since Σ is identifiable from the model. This relationship and related observations are contained in [35].

Example: whitened NGCA model. Let X have a NGCA model as in Definition 1.1.1. If $\Sigma = \mathrm{Cov}(X) = I_p$, then by Proposition 1.2.13 the non-Gaussian and Gaussian subspaces are orthogonal complements in the usual Euclidean inner product. As the next proposition demonstrates, if Σ is not the identity but still positive definite, we can "whiten" (and center) X so that the non-Gaussian and Gaussian spaces are orthogonal:

Proposition 1.2.14. *Let $X \in \mathbb{R}^p$ have a NGCA decomposition as in Definition 1.1.1 Let $\mu = \mathbb{E}(X)$ and suppose $\Sigma = \mathrm{Cov}(X)$ exists and is positive definite. Set $\tilde{X} = \Sigma^{-\frac{1}{2}}(X - \mu)$. Then $\mathbb{E}(\tilde{X}) = 0$, $\mathrm{Cov}(\tilde{X}) = I_p$, and $\tilde{X}$ has an NGCA decomposition with non-Gaussian subspace $\Sigma^{\frac{1}{2}}\,\mathrm{span}(\Gamma)$ orthogonal to the Gaussian subspace $\Sigma^{\frac{1}{2}}\,\mathrm{span}(\eta)$.*

Furthermore, if the NGCA decomposition of X is identifiable, the NGCA decomposition of $\tilde{X}$ is identifiable.

From a statistical viewpoint, the structure of NGCA models suggests whitening the data: either by the population covariance matrix if it is known, or by an empirical estimate if it is unknown. Then estimation of the non-Gaussian subspace can be performed under the assumption that it is orthogonal to the Gaussian space, reducing the complexity of the parameter space. The same pre-processing step is often used in ICA ([27], Ch. 6)

Representation of the density and final version of Stein-like identity. Here we give a useful representation for the density of NGCA models with invertible covariance matrices, and the final form of the Stein-like identity.

Proposition 1.2.15. *Let the p-dimensional random vector X have a NGCA decomposition as in Definition 1.1.1 such that $\Sigma = \mathrm{Cov}(X)$ with $\Sigma \succ 0$. Let the non-Gaussian vector V have a differentiable density and assume the Gaussian component G has zero mean. Then the density $p(x)$ of X has the form*

$$p(x) = q(\Gamma^T x)\phi_\Sigma(x)$$

for some function q differentiable in x and ϕ_Σ the density of the $\mathcal{N}(0,\Sigma)$ distribution.

This representation for the density is proved in [35], Theorem 1. The proof we give uses whitening. This representation yields the final version of the Stein-like identity for NGCA when the covariance is invertible:

Proposition 1.2.16. *Let the p-dimensional random vector X have a NGCA decomposition as in Definition 1.1.1 such that $\Sigma = \mathrm{Cov}(X)$ with $\Sigma \succ 0$. Let the non-Gaussian vector V have a differentiable density and assume the Gaussian component G has mean 0. Then for a real-valued differentiable function g defined on $\mathbb{R}^p$, provided we can differentiate under the integral sign, we have:*

$$\mathbb{E}\left[\nabla g(X)\right] - \Sigma^{-\frac{1}{2}}\mathbb{E}\left[Xg(X)\right] \in \mathrm{span}(\Gamma).$$

Stein's identity for multivariate Gaussian distributions states
$\mathbb{E}\left[\nabla g(X)\right] - \Sigma^{-\frac{1}{2}}\mathbb{E}\left[Xg(X)\right] \in \mathrm{span}(\Gamma) = 0$ if and only if $X \sim \mathcal{N}(0,\Sigma)$. In the presence of non-Gaussian components, whatever is left over must lie in the non-Gaussian space.

1.3 Connection to Sufficient Dimension Reduction

There is a surprising connection between the NGCA model of Definition 1.1.1 and a set of models proposed by Cook ([10][12]) in the context of sufficient dimension reduction (SDR) in regression ([11][1]). We begin by briefly reviewing the ideas behind SDR. We then discuss Cook's specific model, and make the connection to NGCA.

In the regression problem we have observations on the pair (X,Y) where $X \in \mathbb{R}^p$ is a p-dimensional set of predictors and $Y \in \mathbb{R}$ is a scalar response. The goal is to use the observations to make inferences on the conditional distribution of $Y|X$. However in modern data sets p is often large (perhaps larger than the number of observations), so reducing the dimensionality of X is an essential pre-processing step. Many approaches focus on variable selection - the elimination of variables judged to bear no relation to the response Y. For p small this includes subset selection methods like backward and forward selection ([19], Section 13.2.2), or the use of criterion like AIC, BIC, Adjusted R^2 or Mallows' Cp ([19], Section 22.1.1). However, as p increases the computational complexity of these approaches increases combinatorially. A popular alternative approach is to use the LASSO, an ℓ_1-penalized procedure for automatic predictor selection in such high-dimensional cases [39]. Other approaches use the data to derive new covariates, or input features, on which Y is regressed. In principal components regression (PCR), the input features are the principal components of the original predictor variables. In partial least squares (PLS), input features are derived from the data which have large variance and large correlation with Y [22].

SDR provides a general theoretical framework for classifying informative (for the regression problem) dimension-reducing transformations of X. A transformation $R : \mathbb{R}^p \to \mathbb{R}^k$,

$k \leq p$ is a *sufficient dimension reduction* if $Y|X \overset{d}{=} Y|R(X)$. Thus for sufficient dimension reductions R we can replace X by $R(X)$ without losing any information for the regression. For simplicity, the SDR literature focuses on linear sufficient dimension reductions $R(X) = \theta^T X$ for $\theta \in \mathbb{R}^{p \times k}$. Furthermore, the condition $Y|X \overset{d}{=} Y|\theta^T X$ is equivalent to the conditional independence condition $Y \perp\!\!\!\perp X \big| \Pi_{\text{span}(\theta)} X$ ([11]). Hence, the goal in SDR is to estimate the *dimension reduction subspace* span(θ) if it exists. Well known examples of methods which estimate the dimension reduction subspace are Sliced Inverse Regression (SIR) [34] and Sliced Average Variance Estimation [13]. For a recent overview of SDR methods, see [1].

From the viewpoint of SDR, [10] proposed models of increasing generality for the distribution of $X|Y$ (called *inverse regression*). The most general model is:

$$X = \mu + \bar{\Gamma}\nu(Y) + N, \tag{1.4}$$

where $\bar{\Gamma} \in \mathbb{R}^{p \times d}$, ν is an unknown function from $\mathbb{R}$ to $\mathbb{R}^d$, $N \sim \mathcal{N}(0, \Delta)$ for $\Delta \succ 0$, and Y is independent of N. The key feature of Cook's models is the existence of a linear sufficient dimension reduction of the form:

$$Y|X \overset{d}{=} Y|\bar{\Gamma}^T \Delta^{-1} X$$

(Proposition 6 in [10]). Hence, the dimension reduction subspace is $\Delta^{-1}\text{span}(\bar{\Gamma})$. This is the parameter of interest.

By inspection, Cook's model for the predictor vector X is exactly the non-Gaussian signal in Gaussian noise model (Definition 1.2.6) when the distribution of $\nu(Y)$ is non-Gaussian. Furthermore, the dimension reduction subspace $\Delta^{-1}\text{span}(\Gamma)$is precisely the non-Gaussian subspace! Both NGCA and SDR seek projections onto low-dimensional linear subspaces; yet that the subspaces should coincide for this model, and the goals of NGCA and SDR should converge in this way, is still surprising, given that the models are motivated in very different contexts. For instance, there is just no analogous notion of a response variable in the NGCA context. Therefore, NGCA methods, which would only utilize the predictor X for estimation, would in principle automatically work in this inverse regression context. But on the other hand, methods specialized for the regression context that make use of the response variables would not work for NGCA in general.

Lemma 1.2.9 explains why the non-Gaussian space and the sufficient dimension reduction space coincide. An interpretation of that lemma is that the non-Gaussian subspace is "sufficient" for the signal when the latter is treated as a parameter to the non-Gaussian signal in Gaussian noise model (Definition 1.2.6). The same is true in model (1.4) when Y is treated as a parameter: the non-Gaussian subspace captures all the information about Y in the distribution of X. We see this by duplicating the arguments of Lemma 1.2.9 for model (1.4), but conditioning on the response Y. This shows that the conditional distribution of $X\big| \left(\bar{\Gamma}^T \Delta^{-1} X, Y = y\right)$ does not depend on Y, whence we can conclude, by the factorization criterion for sufficiency ([33], p. 35, Theorem 6.5), that the conditional density of $X|Y = y$ has the form:

$$p(x|y) = g(x)q(\bar{\Gamma}^T \Delta^{-1} x, y).$$

Therefore, by Bayes' Theorem, the conditional density of $Y|X = x$ has the form:

$$\begin{aligned} p(y|x) &= \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy} \\ &= \frac{g(x)q(\bar{\Gamma}^T \Delta^{-1} x, y)}{\int g(x)q(\bar{\Gamma}^T \Delta^{-1} x, y)dy} \\ &= \frac{q(\bar{\Gamma}^T \Delta^{-1} x, y)}{\int q(\bar{\Gamma}^T \Delta^{-1} x, y)dy}. \end{aligned}$$

For each value of x the conditional distribution of $Y|X = x$ depends on x only through $\bar{\Gamma}^T \Delta^{-1} x$, which implies $\beta = \Delta^{-1}\bar{\Gamma}$ is a sufficient dimension reduction (see [10] Appendix A.1 for a proof using conditional independence).

In the regression setup assumed by Cook, access to the response values $\{Y_i\}_{i=1}^n$ allows us to work directly on the conditional distribution of $X|Y$, which is Gaussian. Since we can write down the form of the Gaussian distribution, we can perform maximum likelihood to estimate the target subspace. When $\Delta = \sigma^2 I_p$ the actual values of Y_i are not needed to deduce the maximum likelihood estimator, which is just the span of the first d principal directions of X (see section 3.2 of [10]; note that this model justifies principal component regression). However, for more complicated forms of the covariance, the maximum likelihood estimator makes use of the response values. This tilts any comparison of NGCA to Cook's maximum likelihood approach against NGCA: more data, in the form of the response values, is available to learn the space, and the estimation is performed by maximum likelihood, which is asymptotically efficient if the underlying model is correct. Moreover, well-known chi-square goodness of fit tests for the true reduced dimension d, based on asymptotic approximations to the distribution of the log-likelihood ratio, are automatically available, unlike in NGCA. These advantages are simply not available to NGCA methods, which currently have no way of incorporating the response information.

However, Cook's maximum likelihood-based method does bear some similarities to NGCA methods proposed in [5][15][16]. A key step in the NGCA algorithms proposed in these articles is to choose a set of "test functions" which are informative for the non-Gaussian space. At best, reasonable heuristics and good judgment are used to identify potentially informative functions: see Section 1.4 of this chapter for more details. Cook's methodology also requires choosing informative functions based mainly on good judgment. The reason is the following: for covariance matrices Δ more complex than multiples of the identity, when the function ν in (1.4) is completely unknown, the maximum likelihood estimator does not always exist (see [10], Section 6.1). To remedy this issue, a completely parametric form of the model called the principal fitted component (PFC) model is introduced:

$$X = \mu + \bar{\Gamma}\beta f(Y) + N,$$

where $\beta \in \mathbb{R}^{d\times r}$ for $r \geq d$ is an unknown matrix of basis coefficients, and f is a r-dimensional vector of *known* basis functions. For this model, the maximum likelihood estimate of the subspace exists (see [12], Corollary 3.4).

Naturally, the key to estimating the PFC model is to choose suitable functions f. Cook recommends plotting each predictor variable against the response as a guide to make reasonable choices. As long as the chosen functions f are sufficiently correlated with the true function ν, the maximum likelihood estimate is $\sqrt{n}$-consistent ([12], Theorem 3.5).

Cook's approach has the advantage of modeling the functions f naturally. Via these functions, the observations on the response variables $\{Y_i\}_{i=1}^n$ are incorporated into the maximum likelihood estimator, thereby providing more information about the non-Gaussian space. However, beyond eyeballing p separate predictor-response plots, Cook does not propose any criteria for selecting informative functions, or crucially, for screening out uninformative functions. Fitting noisy uninformative functions could lead to a poor estimate. The NGCA methods proposed in [5][15][16], on the other hand, propose data-driven heuristics to determine if functions are informative. Above all, both methodologies require the user to use good judgment to tune the algorithm.

1.4 Review of NGCA methods.

The goal of NGCA is to estimate the non-Gaussian space. Assume we observe $X_1, \dots, X_n$, n i.i.d. copies of X, where X has a NGCA decomposition as in (1.1.1). As far as we know, all methods proposed in the literature assume the covariance $\Sigma = \mathrm{Cov}(X)$ exists with $\Sigma \succ 0$. They also make the assumption (often implicitly) that the non-Gaussian space is identifiable and that the dimension of the non-Gaussian space d is known. There are two classes of NGCA methods in the literature: methods based on **joint matrix diagonalization** and methods based on a version of **Stein's identity** that holds for NGCA models.

Joint matrix diagonalization.

We now summarize joint diagonalization at a high level, following the approach discussed in [30], [31] and [29]. Suppose there exist complex $p \times p$ matrices M_k for $k = 1, \dots, K$ with the following property: for an orthogonal $p \times p$ matrix $\mathcal{O}_0 = [\Gamma_0 \quad \Gamma_{0,\perp}]^T$ such that $\Gamma_0 \in \mathbb{R}^{p\times d}$ is an orthogonal projector on the non-Gaussian space, we can write $\mathcal{O}_0 M_k \mathcal{O}_0^T$ blockwise as:

$$\mathcal{O}_0 M_k \mathcal{O}_0^T = \begin{bmatrix} \Gamma_0^T M_k \Gamma_0 & 0 \\ 0 & 0 \end{bmatrix}, \quad k = 1, \dots, K.$$

Thus $\mathcal{O}_0$ block-diagonalizes the M_k. Given such a set of matrices, we can easily deduce a contrast function which is maximized by projections on the non-Gaussian subspace; it is given in the following proposition:

Proposition 1.4.1. *The criterion $Q(\Gamma)$ defined on the space of $p \times d$ orthogonal matrices by*

$$Q(\Gamma) = \sum_{k=1}^{K} \|\Gamma^T M_k \Gamma\|_F^2,$$

for $\|\cdot\|_F$ *the Frobenius norm is maximized when* $\Gamma = \Gamma_0 U$ *for some* $d \times d$ *orthogonal matrix* U*; that is,* Γ *and* Γ_0 *have the same column space.*

Note: The Frobenius norm is defined on complex matrices M by $\|M\|_F^2 = \mathrm{Tr}(MM^*)$ where M^* is the complex conjugate of M.

Optimizing the criterion presented in Proposition 1.4.1 is accomplished by iteratively solving a generalized eigenvalue problem; see [31] and [29] for details.

To utilize this approach we need the orthogonal matrix $\mathcal{O}_0$ to recover the independent non-Gaussian and Gaussian components as in Definition 1.1.1, so that $\Gamma_0^T X$ is non-Gaussian and independent of the Gaussian random vector $\Gamma_{0_\perp}^T X$. This means that the non-Gaussian and Gaussian spaces must be orthogonal complements. Pre-whitening is an essential preprocessing step in joint diagonalization to ensure orthogonality: we transform our sample by $Y_i = \Sigma^{-\frac{1}{2}} X_i$ and use joint diagonalization to estimate the *whitened non-Gaussian subspace* spanned by the columns of $\Sigma^{\frac{1}{2}} \Gamma_0$. We then "pull back" the estimate by $\Sigma^{-\frac{1}{2}}$ to obtain an estimate of the non-Gaussian space on the original scale. Usually, the covariance Σ is unknown, and we whiten by a consistent estimator of $\Sigma^{-1/2}$. Henceforth, we shall assume that X has orthogonal Gaussian and non-Gaussian spaces.

The only matrices M_k proposed in the NGCA literature with the appropriate diagonalization properties are matrices whose entries are generalized fourth-order cumulants, and the Hessian matrix of the log characteristic function. The fourth order cumulant approach is examined primarily in [30] but also covered in [31] and [29]. Consider the $p \times p$ fourth order cumulant matrices M^{kl}, $k, l = 1, \ldots, p$, defined element-wise by $\left(M^{kl}\right)_{ij} = \mathrm{cum}(X_i, X_j, X_k, X_l)$, where

$$\begin{aligned}\mathrm{cum}(X_i, X_j, X_k, X_l) = {} & \mathbb{E}\left(X_i X_j X_k X_l\right) - \mathbb{E}\left(X_i X_j\right) \mathbb{E}\left(X_k X_l\right) \\ & - \mathbb{E}\left(X_i X_k\right) \mathbb{E}\left(X_j X_l\right) - \mathbb{E}\left(X_i X_l\right) \mathbb{E}\left(X_j X_k\right).\end{aligned}$$

It can be shown that, if $\mathcal{O}_0 = \left[\Gamma_0 \quad \Gamma_{0,\perp}\right]^T$ recovers the independent non-Gaussian and Gaussian components, we have $\left(\mathcal{O}_0 M_{kl} \mathcal{O}_0^T\right)_{ij} = \mathrm{cum}\Big(\left(\mathcal{O}_0 X\right)_i, \left(\mathcal{O}_0 X\right)_j, X_k, X_l\Big)$; and therefore, if $i > d$ or if $j > d$, then $\left(\mathcal{O}_0 M_{kl} \mathcal{O}_0^T\right)_{ij} = 0$. This is precisely the desired diagonalization property. In practice, we do not know the population fourth order cumulants, so we form estimates $\widehat{M}_{kl}$ whose elements contain the empirical fourth order cumulants (replacing the population expectations that define the fourth order cumulants by sample averages). We then maximize $Q(\Gamma) = \sum_{k,l=1}^{p} \|\Gamma^T \widehat{M}_{kl} \Gamma\|_F^2$ over $p \times d$ orthogonal matrices Γ as suggested by Proposition 1.4.1.

The Hessian of the log characteristic function of X is also simultaneously diagonalizable by $\mathcal{O}_0$. This is the approach to NGCA studied primarily in [31] and [29]; we summarize the key ideas now. For a generic random vector Z the characteristic function of Z at t is defined by

$$\mathcal{X}(t; Z) = \mathbb{E}\left(\exp(it^T Z)\right),$$

where i is the imaginary unit: $i^2 = -1$. Assuming X has independent non-Gaussian and Gaussian components recovered by $\mathcal{O}_0$, we can show, using the usual properties of characteristic functions,

$$\mathcal{X}(t; X) = \mathcal{X}(\Gamma_0^T t;\ \Gamma_0^T X) \exp\left(-\frac{1}{2}\|\Gamma_{0_\perp}^T t\|_2^2\right)$$

(recall that we assume X is white: $\mathrm{Cov}(X) = I_p$). Using the chain rule, it is not hard to show that the Hessian of the log of $\mathcal{X}(t; X)$ is equal to

$$\nabla^2 \log \mathcal{X}(t; X) = \mathcal{O}_0^T \begin{bmatrix} \nabla^2 \log \mathcal{X}(\Gamma_0^T t;\ \Gamma_0^T X) & 0 \\ 0 & -I_{p-d} \end{bmatrix} \mathcal{O}_0.$$

Therefore the matrices $M_k = \nabla^2 \log \mathcal{X}(t_k; X) + I_p$ for some collection of p-dimensional vectors t_k are jointly diagonalizable:

$$\mathcal{O}_0 M_k \mathcal{O}_0^T = \begin{bmatrix} \nabla^2 \log \mathcal{X}(\Gamma_0^T t_k;\ \Gamma_0^T X) + I_d & 0 \\ 0 & 0 \end{bmatrix}.$$

To form the sample estimate $\widehat{M_k}$ of M_k, compute the empirical characteristic function $\widehat{\mathcal{X}}(t_k; X) = \frac{1}{n}\sum_{m=1}^n \exp(it_k^T X_m)$ and then compute $\widehat{M_k} = \nabla^2 \log \widehat{\mathcal{X}}(t_k; X) + I_p$.

Numerical simulations have shown that both approaches – using fourth order cumulants or characteristic functions – recover the non-Gaussian subspace. When the non-Gaussian component has lighter tails than a normal distribution, the estimate of the non-Gaussian space based on fourth order cumulants works better than the characteristic function approach. The reverse is true when the non-Gaussian component has heavier tails [29]; this is not surprising, since outliers can have severe effects on the estimates of fourth order cumulants, which are unstable.

While Proposition 1.4.1 is suggestive, there are no published theoretical results that rigorously prove consistency of the joint diagonalization methods. As such, rates of convergence and asymptotic variances are unavailable. In Chapter 2 of this dissertation, we propose a different characteristic function based approach to NGCA, and providing proofs of consistency and $\sqrt{n}$ asymptotic normality. Meanwhile, research on the joint diagonalization approach to NGCA appears to have lapsed; the most state of the art NGCA methods are based on a structural identity that resembles the famous identity of Stein.

Approaches based on a Stein-like identity.

The second class of approaches to NGCA relies on the following identity, which can be related to Stein's famous identity for the multivariate normal distribution:

A Stein-like identity for NGCA. Let X have a NGCA decomposition as in Definition 1.1.1 such that $\Sigma = \mathrm{Cov}(X)$ exists with $\Sigma \succ 0$. Assume that the Gaussian component G has mean 0. Then for a differentiable function $g : \mathbb{R}^p \to \mathbb{R}$, under mild regularity conditions on the distribution of X, we have:

$$\mathbb{E}\left[\nabla g(X)\right] - \Sigma^{-1}\mathbb{E}\left[Xg(X)\right] \in \mathrm{span}(\Gamma), \tag{1.5}$$

i.e. the vector $\beta(g) = \mathbb{E}\left[\nabla g(X)\right] - \Sigma^{-1}\mathbb{E}\left[Xg(X)\right]$ lies in the non-Gaussian space. The usual Stein identity states that $\beta = 0$ if and only if $X \sim \mathcal{N}(0, \Sigma)$. We interpret the identity in the NGCA context as indicating that once the Gaussian noise is subtracted out, what is left must lie in the non-Gaussian space.

Broadly speaking, the approaches to NGCA based on Stein's identity involve two steps [16]:

1. Given a sample $\{X_i\}_{i=1}^n$ consisting of i.i.d. copies of X, for a collection of differentiable functions $\{g_j\}_{j=1}^J$ form candidate vectors $\widehat{\beta}_j$ based on the $X_i's$ which are suitably "close" to $\beta_j = \beta(g_j)$ ($\beta(g_j)$ depends, of course, on the unknown underlying distribution of X). By the Stein-like identity the vectors $\widehat{\beta}_j$ lie approximately on the non-Gaussian space.

2. From the collection $\{\widehat{\beta}_j\}_{j=1}^J$, extract an estimate of the overall d-dimensional non-Gaussian space.

The key parameter for this class of NGCA methods is the choice of functions $\{g_j\}$. They must be selected in such a way that they are informative for the non-Gaussian space. One class we can rule out right away is linear functions. If $g(x) = a^T X$ then, using $\mathbb{E}\left[XX^T\right] = \Sigma + \mu\mu^T$, where $\mu = \mathbb{E}[X]$, we have,

$$\begin{aligned}\mathbb{E}[\nabla g(X)] - \Sigma^{-1}\mathbb{E}[Xg(X)] &= a - \Sigma^{-1}\left(\Sigma + \mu\mu^T\right)a \\ &= (a^T\mu)\Sigma^{-1}\mu.\end{aligned}$$

Therefore, all linear functions just yield scalar multiples of the vector $\Sigma^{-1}\mu$ (this vector lies in the non-Gaussian space due to the fact we assume the Gaussian component G has mean 0). Usually, we assume that $\mu = 0$ (or we empirically center the data vectors), in which case the class of linear functions is completely trivial.

Thus nonlinear functions are preferable to linear functions. Still, the choice of functions remains an important tuning parameter to such algorithms. If too few functions are selected, we may only recover a proper subset of the non-Gaussian space, potentially losing information. If too many functions are selected, there may be a non-negligible fraction of vectors which are uninformative for the non-Gaussian space, resulting in a noisy, high-variance estimates. Currently, the choice of functions is guided by a combination of reasonable choices and data-driven heuristics. On the one hand, the algorithms are extremely flexible for users;

on the other, there do not seem to be any solid theoretical results concerning how to choose functions in a suitable (or optimal) way.

NGCA by the procedure of Blanchard et. al. [5] The first NGCA procedure based on the Stein-like identity (1.5) in the literature was proposed in [5]. We now describe the algorithm and discuss its performance. See Figure 4, p. 259 of that article for the complete pseudocode of the algorithm.

Formation of $\widehat{\beta}_j$*.* Data vectors X_i are pre-whitened by the transformation

$$\tilde{X}_i = \widehat{\Sigma}^{-1}(X_i - \bar{X}),$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T$. Centering by the overall mean $\bar{X}$ ensures the Gaussian component of the data is approximately centered, while "whitening" by the empirical covariance $\widehat{\Sigma}$ justifies using estimates of vectors in the non-Gaussian space of the form

$$\hat{\beta}(g) = \frac{1}{n}\sum_{i=1}^n \left[\nabla g(\tilde{X}_i) - \tilde{X}_i g(\tilde{X}_i)\right],$$

as suggested by (1.5).

Multi-index Projection Pursuit step. Ideas from Projection Pursuit ([26]) are utilized to guide reasonable choices of the index functions g_j. In particular, the g_j are restricted to functions of the form:

$$g_j(x) = h_{a_j}(\omega_j^T x),$$

for $j = 1, \ldots, J$, where $\{h_a\}$ is a family of real-valued functions defined on $\mathbb{R}$ indexed by some parameter a. The h_a act like projection pursuit indices, and the ω_j are chosen such that projections $\omega_j^T \tilde{X}_i$, $i = 1, \ldots, n$ are non-Gaussian. The authors recommend choosing h_a from one of the following three families:

$$h_a^{(1)}(x) = x^3 \exp\left(-\frac{x^2}{2a^2}\right), \qquad \text{(Gauss-Pow3)}$$

$$h_a^{(2)}(x) = \tanh(ax), \qquad \text{(Hyperbolic tangent)}$$

$$h_a^{(3)}(x) = \exp(iax). \qquad \text{(Fourier)}$$

These are exactly the functions used by the FastICA procedure, which searches for projection directions that maximize non-Gaussianity [28]. The projection directions ω_j are chosen by iterating the FastICA algorithm a finite number of times using these functions. By incorporating J different non-Gaussian directions this procedure for generating vectors $\widehat{\beta}_j$, which the authors term "Multi-Index Projection Pursuit," can be sensitive to a variety of departures from normality (e.g. heavy tails or multi-modality). This is not true of traditional Projection Pursuit, which optimizes a single fixed non-Gaussian index. Note that the

FastICA step is not run to convergence, to avoid the possibility that the chosen directions align with the strongest non-Gaussian directions in the data: since the goal is to extract the whole non-Gaussian subspace, we need a rich enough class of vectors $\widehat{\beta}_j$ to pick up weaker non-Gaussian directions in the data.

Extracting the non-Gaussian space. In the subspace extraction step, PCA is run on the vectors $\{\widehat{\beta}_j\}$ and the span of the leading d principal directions is taken as the estimate of the whitened non-Gaussian space. Thus we must have $J \geq d$: eigenvectors corresponding to zero eigenvalues represent directions along which the projections of $\widehat{\beta}_j$ have zero variance, making those directions necessarily redundant and noisy. To recover an estimate of the non-Gaussian space of the original data, we can "pull back" to the original space by pre-multiplying the d directions of the PCA step $\widehat{\Sigma}^{-\frac{1}{2}}$.

The main issue with this approach is that PCA is not scale invariant, and the mapping $g \to \widehat{\beta}(g)$ is linear. Therefore multiplication of g by an arbitrarily large scalar could severely impact the results of PCA. Heuristic arguments are adduced to justify normalization, prior to PCA, of each $\widehat{\beta}_j$ by its sample standard error; this is the square root of

$$s^2(\widehat{\beta}_j) = \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla g_j(\tilde{X}_i) - \tilde{X}_i g_j(\tilde{X}_i) - \widehat{\beta}_j\|_2^2 \right) /n.$$

The norm of the normalized vectors $\widehat{\beta}_j / s(\widehat{\beta}_j)$ can be interpreted as a signal-to-noise ratio; $\widehat{\beta}_j$ is excluded from the PCA step if its signal-to noise ratio is smaller than some user-chosen threshold.

Performance. The article reports a number of simulation results in which NGCA is compared to Projection Pursuit (PP) methods when the underlying model is a NGCA model. In general NGCA outperformed the PP methods. Marked improvement is observed in simulations where the non-Gaussian component contains stochastically dependent light-tailed and heavy-tailed distributions. It is known that the Pow3 index is sensitive to sub-Gaussian (light-tailed) departures from normality, while tanh is sensitive to super-Gaussian departures [28]. By combining different indices, NGCA shows superior performance to these fixed index approaches when departures from normality of both kinds are present in the non-Gaussian components.

Simulations were also performed when the dimension of the data increases. Although NGCA outperformed the methods to which it was compared, its performance deteriorated rapidly as dimension increased, and usually returned poor results for dimensions $p \geq 40$. This may be due to the poor performance of the FastICA procedure finding good candidate directions in high-dimensions. The performance of NGCA also deteriorates in simulations when the Gaussian components have an ill-conditioned covariance structure for the same reason. The authors attribute this to the empirical pre-whitening step, where the sample covariance matrix is a poor estimator of the population covariance when the latter is ill-conditioned.

Discussion. The article [5] is a foundational paper in NGCA. The heuristics seem convincing, and the simulations persuasive that it is a useful method for learning non-Gaussian structure. It also represents an advance in theoretical sophistication over the joint diagonalization approaches detailed in [30][31][29]. Theorems 3 and 4 in particular give some justification for the algorithm. We summarize the theorems briefly. Let $\mathcal{S}$ be a linear subspace and define the distance from a vector on $\mathbb{R}^p$ to $\mathcal{S}$ in the usual way:

$$dist(\beta, \mathcal{S}) = \inf_{\gamma \in \mathcal{S}} \|\beta - \gamma\|_2.$$

In Theorem 3, we assume the set of estimating vectors $\widehat{\beta}_1, \ldots, \widehat{\beta}_J$ defined by

$$\widehat{\beta}_j = \widehat{\beta}(g_j) = \frac{1}{n} \sum_{i=1}^{n} \left[\nabla g_j(\tilde{X}_i) - \tilde{X}_i g_j(\tilde{X}_i) \right]$$

are computed using data $\tilde{X}_i$ whitened by the true population covariance matrix. If the columns of $\Gamma \in \mathbb{R}^{p \times d}$ span the true non-Gaussian space, then with high probability,

$$dist(\widehat{\beta}_j, \operatorname{span}(\Gamma)) \leq \sqrt{\frac{\log J + \log d}{n}},$$

uniformly in $j = 1, \ldots, J$. Theorem 4 concludes that, under some additional assumptions, when the data are whitened by the sample covariance matrix we have:

$$dist(\widehat{\beta}_j, \operatorname{span}(\Gamma)) \leq \sqrt{\frac{d \log n}{n}} + \sqrt{\frac{\log J}{n}},$$

uniformly in $j = 1, \ldots, J$. In some ways, however, the theory as it appears in the article is unsatisfying:

1. Both theorems make the assumption that there exist a constant λ such that

   $$\mathbb{E}\left[\exp\left(\lambda \|X\|_2^2\right)\right] < \infty,$$

   which implies that every moment of X is finite. The authors themselves are careful to note that this excludes some super-Gaussian distributions that may be of practical interest. By contrast, the characteristic function based method for NGCA we propose in Chapter 2 only assumes the non-Gaussian distribution has finite fourth moments.

2. Theorem 4 could probably be refined to eliminate the $\sqrt{\log n}$ term This would show the method achieves the parametric rate $n^{-1/2}$ even under empirical pre-whitening. The NGCA method we propose in Chapter 2 attains $\sqrt{n}$-consistency under empirical pre-whitening.

Another deficiency in the theory developed in the article is that it does not guarantee consistency or provide convergence rates for the estimated non-Gaussian space itself. To illustrate, if the columns of $\widehat{\Gamma}_n$ span the estimated non-Gaussian subspace, there is no theorem which bounds the deviation

$$\frac{1}{2d}\|\Pi_{\text{span}(\widehat{\Gamma}_n)} - \Pi_{\text{span}(\Gamma)}\|_F^2,$$

which is the error criterion used in the paper. The problem appears to be that, while for any collection of candidate functions $g_1, \ldots, g_j$ we can bound uniformly the distance between $\widehat{\beta}_j$ and the non-Gaussian space, it is difficult to guarantee that the vectors $\widehat{\beta}_j$ are rich enough so that the full non-Gaussian space can be recovered by PCA. For instance, weak non-Gaussian signals may not be detected. By contrast, the theoretical results for the method presented in Chapter 2 demonstrate the consistency and asymptotic normality of estimates of the non-Gaussian space as a whole.

We do not however wish to convey too pessimistic an impression of this work: the simulation results displayed in [5] are clear evidence of the efficacy of the procedure, and subsequent NGCA methods based on this algorithm address the problem of accurate estimation of the space itself [15][16].

NGCA with radial kernel functions. Another algorithm for NGCA is proposed in [32]. The principal difference of this algorithm from the algorithm outlined in [5] is that functions g_j are not chosen according to the multi-index Projection Pursuit method. Instead, the algorithm is motivated by the following observations. Let $\{\tilde{X}_i\}_{i=1}^n$ be the whitened data points. Heuristic arguments provided in [5] indicate that vectors $\widehat{\beta}(g) = \frac{1}{n}\sum_{i=1}^n \left[\nabla g(\tilde{X}_i) - \tilde{X}_i g(\tilde{X}_i)\right]$ which are informative for the non-Gaussian space have a large norm relative to their estimated standard error $s(\widehat{\beta}(g))$, where

$$s^2(\widehat{\beta}(g)) = \left(\frac{1}{n}\sum_{i=1}^n \|\nabla g(\tilde{X}_i) - \tilde{X}_i g(\tilde{X}_i) - \widehat{\beta}(g)\|_2^2\right)/n.$$

Therefore, a reasonable method to select functions g is to restrict g to lie in some parameterized family of functions g_θ, and then to optimize $\|\widehat{\beta}(g_\theta)\|_2^2/s^2(\widehat{\beta}(g_\theta))$ over θ. The authors propose:

$$g_{\sigma,M,a}(x) = \sum_{i=1}^n a_i K_{\sigma,M}(x, \tilde{X}_i),$$

where $K_{\sigma,M}(x,y) = \exp\left\{-\frac{1}{2\sigma^2}(x-y)^T M(x-y)\right\}$ is a Gaussian radial kernel function with $M \succ 0$. Functions such as $g_{\sigma,M,a}$ are often found in regularized function estimation problems in machine learning (see e.g. [22], Section 5.8.1). Some algebra yields

$$\frac{\|\widehat{\beta}(g_{\sigma,M,a})\|_2^2}{s^2(\widehat{\beta}(g_{\sigma,M,a}))} = \frac{a^T F a}{a^T G a}$$

for $n \times n$ matrices F and G which depend on σ^2, M and the data observations $\tilde{X}_1, \ldots, \tilde{X}_n$ (see Equations (11) and (12) in [32]). For fixed values of σ^2 and M we can optimize this criterion over the weight vector a; the well-known solution to this generalized eigenvalue problem is the eigenvector of $G^{-1}F$ with largest eigenvalue.

We first describe a single run of the algorithm. Choose values of σ^2 along some preselected grid to obtain $\sigma_1^2, \ldots, \sigma_J^2$. Then for a fixed value of M, solve the generalized eigenvalue problem to obtain the weight vector a_j, $j = 1, \ldots, J$, and output $\widehat{\beta}_j = \widehat{\beta}(g_{\sigma_j^2, M, a_j})$. PCA is then used to compute an estimate of the non-Gaussian space based on $\{\widehat{\beta}_j\}$.

In order to improve the estimates, the algorithm is iterated. The iterative procedure is called Iterative Metric Adaptation for Radial Kernel Functions (IMAK). At each iteration, the matrix M defining the norm of the kernel is chosen in an adaptive fashion. At time $t = 0$, set $M_0 = I_p$ and output vector estimates $\widehat{\beta}_j^{(0)}$. At time t, $t = 1, 2, \ldots$, set:

$$M_t = \sum_{j=1}^{J} \widehat{\beta}_j^{(t-1)} \left(\widehat{\beta}_j^{(t-1)} \right)^T .$$

Rescale to make the trace of M_t equal to d for all iterations. The idea of the adaptation step is that the major axes of the level sets of the kernels $K_{\sigma_j^2, M_t, a_j}$, which are ellipsoids, will lie in Gaussian directions, meaning that the kernels will change more rapidly in non-Gaussian directions. At each step of the algorithm, greater sensitivity to the non-Gaussian directions is (hopefully) achieved. To guard against the possibility that a small number of strong non-Gaussian directions will dominate weaker signals, the authors propose another method for updating M_t: set

$$M_t = \bar{\lambda} \sum_{k=1}^{d} u_k u_k^T + \sum_{k=1}^{J} \lambda_k u_k u_k^T,$$

where $\lambda_1, \ldots, \lambda_J$ are the eigenvalues of $\sum_{j=1}^{J} \widehat{\beta}_j^{(t)} \left(\widehat{\beta}_j^{(t)} \right)^T$ with eigenvectors $u_1, \ldots, u_J$, and $\bar{\lambda} = \frac{1}{d} \sum_{k=1}^{d} \lambda_k$; re-scale M_t to have trace equal to d. This ensures equal weights on all non-Gaussian directions.

Simulation results show the procedure with the iterative IMAK step is comparable to NGCA with Multi-Index Projection Pursuit, and in some cases has lower error. This indicates that adapting to the underlying non-Gaussian structure in a data-driven fashion can improve performance of NGCA algorithms. While there is no rigorous theory in the article to buttress this conclusion, the adaptation idea appears in more recent NGCA algorithms, anecdotal evidence of its worth.

Sparse Non-Gaussian Components Analysis (SNGCA) The procedure put forth in [15] called Sparse Non-Gaussian Component Analysis (SNGCA) is a variant of the Multi-index Projection Pursuit procedure of [5]. Along with some minor differences, the SNGCA procedure differs in two key respects:

1. SNGCA is adaptive: the algorithm can be iterated to improve the the quality of estimates of the non-Gaussian space and learn new directions.

2. Rather than estimating the non-Gaussian space by PCA, SNGCA estimates it by the span of the major axes of a certain ellipsoid containing the convex hull of the vectors $\widehat{\beta}_j$.

The reason for performing adaptation–if it works–is obvious. The reason for computing the rounding ellipsoids for subspace extraction is, according to the authors, because PCA can be noisy when there are many candidate vectors $\widehat{\beta}_j$, $j = 1, \dots, J$, such that a nontrivial fraction are uninformative or lie close to the Gaussian space. The estimation error of the bounding ellipsoid approach does not significantly increase with the number of candidate vectors: in Theorem 3, the error of the whole estimated non-Gaussian space is bounded up to constants by the maximum distance of the $\widehat{\beta}_j$ to the target space, with no dependence on J. The kind of theoretical results offered in this article are an improvement over those available in [5]. We now describe the SNGCA algorithm.

Formation of candidate vectors $\widehat{\beta}_j$. SNGCA avoids the problems associated with whitening by the sample covariance matrix by choosing functions g such that $\frac{1}{n}\sum_{i=1}^{n} X_i g(X_i) = 0$. Then by the Stein identity (1.5) the quantity $\frac{1}{n}\sum_{i=1}^{n} \nabla g(X_i)$ should lie approximately in the non-Gaussian space. The authors recommend scaling each coordinate of the X_i to have variance 1, which requires computation only of the diagonals of the sample covariance.

SNGCA exploits the linearity of the mapping $\widehat{\beta}(g, \Sigma) = \frac{1}{n}\sum_{i=1}^{n} [\nabla g(X_i) - \Sigma^{-1} X_i g(X_i)]$ to find suitable linear combinations of functions. For each $j = 1, \dots, J$ suppose there are K_j fixed functions g_{jk} and a weight vector $c_j \in \mathbb{R}^{K_j}$. Let $g_j(x) = \sum_{k=1}^{K_j} c_{jk} g_{jk}(x)$ for each $j = 1, \dots, J$. Set $\hat{\gamma}(c_j)$ and $\hat{\theta}(c_j)$ as follows:

$$\hat{\gamma}(c_j) = \frac{1}{n}\sum_{i=1}^{n} X_i g_j(X_i)$$
$$\hat{\theta}(c_j) = \frac{1}{n}\sum_{i=1}^{n} \nabla g_j(X_i).$$

Now choose the weight vectors c_j by solving a convex projection problem. For a fixed *probe vector* $\xi \in \mathbb{R}^p$ solve:

$$\begin{aligned} \underset{c \in \mathbb{R}^{K_j} : \|c\|_1 \le 1}{\operatorname{argmin}} \quad & \|\xi - \hat{\theta}(c)\|_2 \\ \text{subject to} \quad & \hat{\gamma}(c) = 0. \end{aligned}$$

Let the outputted set of weights be $\hat{c}_j$. Then it is easy to see $\widehat{\beta}_j = \widehat{\beta}(g_j, \Sigma) = \hat{\theta}(\hat{c}_j)$ for any choice of Σ, rendering whitening unnecessary. Some remarks:

1. The functions g_{jk} are, in general, projection pursuit functions as in Multi-Index Projection pursuit: $g_{jk}(x) = h(x, \omega_{jk})$ for $\omega_{jk} \in \mathbb{R}^p$.

2. The ℓ_1 penalty in the convex projection problem helps to bound estimation error (see Theorem 2) and also outputs weight vectors $\hat{c}_j$ that are sparse, i.e. that have entries equal to 0. Therefore, the procedure is a data-driven way to eliminate possibly noisy functions g_{jk}.

3. The probe vectors ξ_j and projection directions ω_{jk} yield good candidate vectors $\widehat{\beta}_j$ if they lie in the vicinity of the non-Gaussian space. To build up informative directions, an adaptive procedure is used. At time $t = 0$ we sample the ξ_j and ω_{jk} uniformly on the unit sphere. For each new iteration, a decreasing fraction of the directions are still sampled uniformly, while an increasing fraction are constructed from the estimated non-Gaussian directions found in the previous iteration by taking linear combinations with randomly selected weights. The idea is to choose more informative directions for each round of the procedure, while still guarding against the possibility of each vector converging on the strongest non-Gaussian directions and missing weaker directions. Further details are contained in the article in Algorithm 4 of Appendix C, p. 3045.

Extraction of the non-Gaussian space. Given a $p \times p$ symmetric positive definite matrix B define the ellipsoid $\mathcal{E}_r(B)$ by:

$$\mathcal{E}_r(B) = \{x \in \mathbb{R}^p | x^T Bx \leq r^2\}.$$

Let S be the convex envelope of the candidate vectors $\pm\widehat{\beta}_j$. Then there exists B such that

$$\mathcal{E}_1(B) \subseteq S \subseteq \mathcal{E}_{p^{1/2}}(B).$$

The ellipsoid $\mathcal{E}_{p^{1/2}}(B)$ is called the $\sqrt{p}$-*rounding* ellipsoid for S.

To extract the non-Gaussian space from the collection of vectors $\widehat{\beta}_1, \ldots, \widehat{\beta}_J$ we compute the matrix B defining the $\sqrt{p}$-minimum rounding ellipsoid. We then take the eigendecomposition of B, using the eigenvectors associated with the largest eigenvalues as basis vectors for the estimated non-Gaussian space. As an extra precaution, the authors suggest to test each candidate basis vector for non-normality: project each data point on the candidate basis vector and run statistical tests which are sensitive to departures from the Gaussian distribution. See Appendix A, p. 3042 for more details.

Performance: Results are mixed. While SNGCA is comparable to NGCA (that is, NGCA with Multi-Index Projection Pursuit) on all simulated data, it only shows clear improvement when the non-Gaussian component has thin tails, even if the projection pursuit function is sensitive to heavy tails. NGCA outperforms SNGCA for Gaussian mixtures and in the case the non-Gaussian component has sub-Gaussian and super-Gaussian tails. The results illustrate how sensitive the performance of SNGCA is to the adaptive scheme used to learn the non-Gaussian directions. Adaptation occurs swiftly for thin-tails. But in the Gaussian mixture case, most random projections have a Gaussian distribution. Since the algorithm begins with purely-random projections, the decrease in estimation error is small for each iteration.

The same phenomenon occurs when there are sub- and super-Gaussian components: adaptation occurs slowly. The FastICA-type procedure used to select directions in NGCA seems to do a better job at detecting these forms of non-Gaussianity without iterating the entire estimation procedure. Consequently, some prior knowledge of the type of non-Gaussianity of the data, combined with good heuristics, might be necessary to use SNGCA effectively.

Mixed results are reported as the dimension of the data p is increased. SNGCA provides marked improvement over NGCA when the non-Gaussian components are thin-tailed, but is comparable to (or worse) than NGCA when there are other forms of non-normality in the data. In these situations, the structural adaptation step does not perform much better than the FastICA-type procedure in NGCA in choosing promising non-Gaussian directions.

We do however see a very clear improvement of SNGCA over NGCA when the covariance structure of the Gaussian components is ill-conditioned. This is mostly likely due to the fact that this algorithm avoids estimation of the sample covariance.

Discussion: The authors call this NGCA method "sparse" and in the conclusion they claim the method provides an estimate of the true non-Gaussian dimension d. However these claims are never clearly explained. Presumably, SNGCA is sparse because the ℓ_1 penalty on the weight functions c_j returns solutions with many entries equal to 0. In the conclusion, the authors state "SNGCA provides an estimate for the dimension of the non-Gaussian subspace." We assume this estimate of the true dimension is obtained when the data projected on the principal axes of the minimum rounding ellipse are subject to tests of non-normality. All directions for which we cannot reject the null hypotheses are not included in the basis of the estimated non-Gaussian space, delivering an estimate of the true dimension. However, the paper does not explain this clearly, nor does it explain how the significance levels of these tests should be set, whether the user should correct for multiple comparisons, etc. In both respects – sparsity and estimating the true non-Gaussian dimension – SNGCA seems promising, but due to lack of explanations their practical consequences remain somewhat vague.

The theoretical results in this paper are improved from [5] and more useful. Theorem 2 provides uniform bounds on the distance between the vectors $\widehat{\beta}_j$ (picked according to convex projection) and the true non-Gaussian space of the form:

$$dist(\widehat{\beta}_j, \text{span}(\Gamma)) \leq K\sqrt{\frac{d}{n}}.$$

Here the columns of Γ are a basis for the non-Gaussian space, and K is a constant that depends on the underlying distribution of the data. We therefore achieve the parametric convergence rate $n^{-1/2}$ with no $\log n$ term. Furthermore, the assumptions require only that $\Sigma = \text{Cov}(X)$ exist such that $\Sigma \succ 0$ (much less restrictive than the assumptions of Theorem 4 in [5], which require moments of all orders to exist) along with some mild boundedness assumptions on the index functions g.

The paper also provides bounds on the estimation error of the whole space in Theorem 3. Assuming there exist vectors $\beta_1, \ldots, \beta_J$ in the non-Gaussian space such that

$$\max_j \|\widehat{\beta}_j - \beta_j\|_2 \leq \delta,$$

we have,

$$\|\hat{\Pi} - \Pi^*\|_F^2 \leq \frac{4\delta^2 p\sqrt{p}}{\lambda^* - 2\delta^2},$$

where $\hat{\Pi}$ is the projection operator of the estimated non-Gaussian space, Π^* is the projection operator on the true non-Gaussian space, and λ^* satisfies, by assumption:

$$\lambda_d \left(\sum_{j=1}^{J} \mu_j \beta_j \beta_j^T \right) \geq \lambda^* > 2\delta^2.$$

Here, $\lambda_d(A)$ represents the dth largest eigenvalue of the symmetric non-negative definite matrix A, and μ_j are weights such that $\sum_{j=1}^{J} \mu_j = 1$. It is not clear why the authors do not combine Theorems 2 and 3 via a union bound to obtain $\sqrt{n}$-convergence of $\hat{\Pi}$ in probability, perhaps paying a mild $\log J$ term for the total number of functions. Of course, all these results hold for one iteration of the algorithm: presumably, finer results could be obtained if the adaptation step were included in the analysis. But such an analysis could be challenging.

Theorem 3 captures the heart of the problem of estimating the non-Gaussian subspace from individual vectors which lie close to the space, particularly in the assumption

$$\lambda_d \left(\sum_{j=1}^{J} \mu_j \beta_j \beta_j^T \right) \geq \lambda^* > 2\delta^2.$$

This is a kind of identifiability assumption which ensures that the set of candidate vectors $\widehat{\beta}_j$ is rich enough to capture the whole non-Gaussian space. The main problem with this assumption, however, is that it cannot be verified. Practically, this means we must still rely on reason and good heuristics to choose suitable functions for detecting the non-Gaussian structure in the data.

Sparse NGCA by Semidefinite Programming. The NGCA method outlined in [16], which is called Sparse Non-Gaussian Component Analysis by Semidefinite Programming (SNGCA-SDP), demonstrates superior performance to all other NGCA algorithms in the literature. Simulation results demonstrate the method is superior to NGCA via Multi-Index Projection Pursuit against a variety of departures from normality. The method is also somewhat robust against high dimensions and ill-conditioning of population covariance matrix. Complementing the convincing simulations, the theoretical development in [16] is more complete than its predecessors, and demonstrates convergence of the estimated non-Gaussian subspace to the true non-Gaussian subspace at $\sqrt{n}$-rate.

Constructing an optimization problem. The principal difference between SNGCA-SDP and previous NGCA algorithms based on the Stein identity is that in SNGCA-SDP an optimization problem is formulated, designed to infer the projection matrix Π^* corresponding to the true non-Gaussian space directly. For X a generic NGCA random vector and indices $j = 1, \ldots, J$ consider $p \times J$ matrices U and G defined by:

$$\begin{aligned} U &= \left[\mathbb{E}\left(\nabla g_1(X)\right), \ldots, \mathbb{E}\left(\nabla g_J(X)\right)\right] \\ G &= \left[\mathbb{E}\left(X g_1(X)\right), \ldots, \mathbb{E}\left(X g_J(X)\right)\right]. \end{aligned}$$

By the Stein-like identity (1.5) if there exists a vector $c \in \mathbb{R}^J$ such that $Gc = 0$, then the vector Uc must lie in the non-Gaussian subspace. This implies the projection operator on the non-Gaussian space is the optimizer of the following min-max problem:

$$\begin{aligned} \underset{\Pi}{\text{argmin}} \ \max_{c \in \mathbb{R}^J} \ & \|\left(I_p - \Pi\right) Uc\|_2^2 \\ \text{subject to } \ & \Pi \text{ is a projection matrix on a } d\text{-dimensional subspace of } \mathbb{R}^p \\ & Gc = 0 \end{aligned} \tag{1.6}$$

The optimization problem (1.6) is mimicked to formulate an optimization problem on the observed data, from which an estimate of the non-Gaussian subspace is obtained. To start building up the new optimization problem, define $p \times J$ matrices $\hat{U}$ and $\hat{G}$ by:

$$\begin{aligned} \hat{U} &= \left[\frac{1}{n}\sum_{i=1}^{n} \nabla g_1(X_i), \ldots, \frac{1}{n}\sum_{i=1}^{n} \nabla g_J(X_i)\right] \\ \hat{G} &= \left[\frac{1}{n}\sum_{i=1}^{n} X_i g_1(X_i), \ldots, \frac{1}{n}\sum_{i=1}^{n} X_i g_J(X_i)\right]. \end{aligned}$$

Note the added constraint $\|c\|_1 \leq 1$ plays the role of controlling the estimation error. Suppose with high probability, uniformly in j, we have

$$\|\frac{1}{n}\sum_{i=1}^{n} \nabla g_j(X_i) - \mathbb{E}\left[\nabla g_j(X)\right]\|_2 \leq \rho_n$$

and

$$\|\frac{1}{n}\sum_{i=1}^{n} X_i g_j(X_i) - \mathbb{E}\left[X g_j(X)\right]\|_2 \leq \nu_n.$$

If $\|c\|_1 \leq 1$ it follows

$$\|(\hat{U} - U)c\|_2 \leq \rho_n,$$

and

$$\|(\hat{G} - G)c\|_2 \leq \nu_n.$$

It turns out that we do have the requisite uniform control over the deviations for $\rho_n = O(\sqrt{\min(p, \log J)/n})$ and $\delta_n = O(\sqrt{\min(p, \log J)/n})$.

These observations justify replacing U and G by $\hat{U}$ and $\hat{G}$ in the optimization problem (1.6). Other modifications are made to make the problem easier to solve. Rewrite the objective function in (1.6) as:

$$\|(I_p - \Pi)\hat{U}c\|_2^2 = \mathrm{Tr}(U^T\,(I_p - \Pi)\,\hat{U}cc^T).$$

The objective function resembles a semidefinite program with matrix variable $X = cc^T$. The constraint $\|c\|_1 \leq 1$ becomes the constraint $\sum_{j,k=1}^{J} |X_{jk}| \leq 1$, which is convex. The quantity $\|\hat{G}c\|_2$ can be re-written as $\mathrm{Tr}(\hat{G}X\hat{G}^T)$. However, X as defined has rank 1, which is a non-convex constraint. To remedy this difficulty, the constraint is simply dropped, and the optimization is performed over the space of symmetric positive semidefinite matrices.

The other difficult constraint is that Π is a d-dimensional projection matrix. Thus Π has rank d, with $\mathrm{Tr}(\Pi) = d$ and $I_p \succeq \Pi \succeq 0$. The latter two constraints are convex, but the rank constraint is not. Once again this constraint is simply removed. This yields the final optimization problem:

$$\begin{aligned} \min_{\mathcal{P}} \max_{X} \quad & \mathrm{Tr}\left(\hat{U}^T(I - \mathcal{P})\hat{U}X\right) \\ \text{subject to} \quad & I_p \succeq \mathcal{P} \succeq 0 \\ & \mathrm{Tr}(\mathcal{P}) = d \\ & X \succeq 0 \\ & \sum_{j,k=1}^{J} |X_{jk}| \leq 1 \\ & \mathrm{Tr}(\hat{G}X\hat{G}^T) \leq \delta^2, \end{aligned} \tag{1.7}$$

where we have used a slack variable δ^2 instead of 0 to help ensure that the optimal c^* in (1.6) is feasible for (1.7) as $X^* = c^*(c^*)^T$ (note that, since $Gc^* = 0$, this will hold with high probability if $\delta^2 \geq \nu_n$).

Having solved (1.7) and obtained the optimum $\hat{\mathcal{P}}$, we compute the projector $\hat{\Pi}$ on the estimated non-Gaussian space from the span of d principal eigenvectors of $\hat{\mathcal{P}}$.

The optimization problem (1.7) is a saddlepoint problem on the domain of positive semidefinite matrices with convex constraints. State of the art algorithms are necessary for solving it. According to the authors, the main drawback of the procedure is that its implementation is computationally demanding. For more details on how to solve the problem, see Section 4 of the article.

Structural adaptation. Structural adaptation is used in SNGCA-SDP to obtain more informative functions g_j at each iteration of the algorithm. Just as in Multi-Index Projection

Pursuit NGCA and SNGCA, functions g are restricted to the form $g(x) = h(x, \omega)$ where h is a FastICA or projection pursuit function such as tanh. At time $t = 0$ the directions $\omega_1, \ldots, \omega_J$ are all sampled uniformly on the unit sphere. At time $t = 1, 2, \ldots$, a fraction of ω_j are drawn from a $\mathcal{N}(0, \hat{\Pi}^{(t-1)})$ distribution, where $\hat{\Pi}^{(t-1)}$ is the projector of the estimated non-Gaussian space computed at time $t - 1$. The remaining fraction are sampled uniformly on the unit sphere; this helps prevent all of the chosen directions from converging on the strongest non-Gaussian signals early in the algorithm, allowing for weaker non-Gaussian signals to be detected.

Performance. SNGCA-SDP is the state of the art NGCA algorithm. It outperforms NGCA with Multi-index Projection Pursuit in every simulation design explored in the paper. The method performs well in moderate dimensions (around $p = 50$). Since the procedure doesn't require pre-whitening, it is robust to ill-conditioned Gaussian noise components like SNGCA. The main drawback, by the authors' own admission, is that implementing solvers for the optimization problem (1.7) is difficult.

Discussion. The theoretical development is the most sophisticated and complete of all the NGCA papers. Besides some mild assumptions on the underlying distribution and the projection indices h, the main assumption is that the projector on the true non-Gaussian space Π^* is in some sense identifiable. The assumption (Assumption 1, p. 218) is as follows: given the collection of functions $g_1, \ldots, g_J$, there exist vectors $c_1, \ldots, c_m$ with $d \leq m \leq J$ such that $Gc_k = 0$, $k = 1, \ldots, m$, and there exist constants $\mu^1, \ldots, \mu^m$ such that

$$\Pi^* \preceq \sum_{k=1}^{m} \mu^k U c_k c_k^T U^T.$$

In other words, the collection of functions $g_1, \ldots, g_j$ is rich enough such that the null space of G spans a sufficiently large subset of the non-Gaussian space. Under this (uncheckable) assumption Theorem 1 of [16] guarantees that with high probability,

$$\|\hat{\Pi} - \Pi^*\|_F^2 \leq C\mu^* \frac{\min(p, \log J)}{n},$$

where C is a constant and $\mu^* = \mu^1 + \ldots + \mu^m$. This shows the full non-Gaussian space can be recovered at $\sqrt{n}$ rate with a mild logarithmic penalty as long as p stays fixed.

Conclusion.

The joint diagonalization algorithms proposed in [30] [31][29] are elegant and are relatively easy to implement. But theoretical development for these algorithms is lacking. In Chapter 2 of this dissertation, we propose a different characteristic based method for NGCA, and provide rigorous proofs of consistency and $\sqrt{n}$ asymptotic normality under mild assumptions.

Methods based on the Stein-like identity (1.5) have taken prominence in the field of late. The state of the art NGCA algorithm is SNGCA-SDP [16]. It performs well under a variety of departures against normality and in moderately large dimensions. However, numerically, it is difficult to implement. If accuracy is to be traded for computational simplicity, SNGCA [15] does not appear to be a good compromise: it only appreciably outperforms NGCA with Multi-index Projection Pursuit [5] under certain kinds of departures from normality. For other kinds of departures, the gains are minimal, and the convergence can be slow, taking many iterations for the algorithm to adapt to the non-Gaussian space. NGCA with Multi-index Projection Pursuit may be a better compromise of accuracy for simplicity.

A crucial tuning parameter to every NGCA method based on the Stein-like identity is how to choose the test functions for finding the non-Gaussian space. While good data-driven heuristics for choosing functions are offered, we do not know of any theoretical guarantees that the functions chosen are rich enough to recover the whole non-Gaussian space. Often this is simply imposed by assumption ([15], [16]). The NGCA method outlined in Chapter 2 of this dissertation avoids this problem by estimating the characteristic function of the data: test functions do not need to be chosen. We provide theoretical guarantees that the method recovers the whole non-Gaussian subspace at the $\sqrt{n}$-rate under mild assumptions. However, unlike [15] and [16] the method does require empirical pre-whitening of the data, which can harm estimates in high-dimensional settings where covariance matrix estimation is difficult. Further theoretical work and simulation studies would shed more light on the comparative advantages of each method.

Chapter 2

NGCA by a Characteristic Function Approach

2.1 Introduction

In this chapter we propose and analyze a NGCA method based on the characteristic function. The use of characteristic functions for NGCA was previously explored in [31]. Their method exploits the fact that when the data are projected onto the correct non-Gaussian and Gaussian spaces, the Hessian of the logarithm of the characteristic function is blockwise diagonal. They propose estimating the non-Gaussian space by finding a projection that simultaneously diagonalizes empirical estimates of the Hessian evaluated at a given finite number of points. A small simulation study is included, but no theoretical guarantees for the performance of the estimator are given.

The method we analyze in this chapter is based on comparing characteristic functions to detect independent components. We call this method the characteristic function-based NGCA estimate, or CHFNGCA. It is adapted from a method first proposed in the context of Independent Components Analysis in [18] and studied in depth in [7]. We provide theoretical guarantees for the performance of the proposed estimator, including consistency and $\sqrt{n}$-asymptotic normality under mild conditions.

The chapter is organized as follows: in Section 2 we propose a characteristic function-based estimator and prove some basic results. Section 3 contains theorems which show that the method is consistent and asymptotically normal. The appendix contains detailed proofs along with some supplementary material.

2.2 The estimator

In this section we review how the NGCA model interacts with the properties of the characteristic function, motivating CHFNGCA. We review the NGCA model and state some key assumptions pertaining to it. We then review the characteristic function, and show how

CHFNGCA is a sensible estimate of the non-Gaussian space. Finally, we prove some basic results about CHFNGCA.

2.2.1 The model.

We remind the reader of the NGCA decomposition first introduced in Chapter 1:

Definition 2.2.1 (NGCA Decomposition.)**.** *We say a p-dimensional random vector X has a d-dimensional NGCA decomposition ($d < p$) if there exists a $p \times d$ matrix Γ and a $p \times (p-d)$ matrix η such that:*

$$\begin{bmatrix} \Gamma^T X \\ \eta^T X \end{bmatrix} = \begin{bmatrix} V \\ G \end{bmatrix} \tag{NGCA Decomposition}$$

where the random vector $V \in \mathbb{R}^d$ has a non-Gaussian distribution, independent of the $(p-d)$-dimensional Gaussian vector G.

We assume that we observe n i.i.d. copies the data vector X. The goal is to estimate the subspace generated by the columns of Γ. This space is the *non-Gaussian subspace*. Projection of the data vectors along this subspace conserves the interesting non-Gaussian structure and eliminates the uninteresting Gaussian structure. The subspace spanned by the columns of η is called the *Gaussian subspace*.

These are the **key assumptions** we make about the model in Definition 2.2.1 throughout the chapter:

1. $\Sigma = \text{Cov}(X)$ exists and $\Sigma \succ 0$. This implies Σ^{-1} exists.

2. The dimension of the non-Gaussian space d is known.

3. The non-Gaussian component V in (2.2.1) does not itself have a d'-dimensional NGCA decomposition for $0 \leq d' < d$ (the case $d = 0$ means V is Gaussian). This implies that the non-Gaussian space is identifiable from the model (see Theorem 1.2.12 in Chapter 1).

2.2.2 Characteristic Functions and the NGCA model.

The characteristic function of a NGCA model given in Definition 2.2.1 possesses a particular structure. In this section, we derive this structure, and demonstrate how it gives rise to a characteristic function-based NGCA estimate we call CHFNGCA. We begin by reviewing some basic properties of the characteristic function.

The characteristic function of a p-dimensional random vector $X \sim P$ is given by:

$$\mathcal{X}(t; P) = \mathbb{E}_P \left[\exp(it^T X)\right].$$

Occasionally we write the characteristic function in terms of random variables: if $X \sim P$ we may use $\mathcal{X}(t;X)$ for $\mathcal{X}(t;P)$. Of interest are these well-known properties of $\mathcal{X}(t;P)$ (see [4], chapter 5, sections 26 and 29):

(i) $\mathcal{X}(t,P)$ exists for every distribution P.

(ii) $|\mathcal{X}(t;P)| \leq 1$.

(iii) The mapping $t \to \mathcal{X}(t;P)$ is uniformly continuous.

(iv) For $X \perp\!\!\!\perp Y$, $\mathcal{X}(t;X+Y) = \mathbb{E}\left[\exp\{it^T(X+Y)\}\right] = \mathbb{E}\left[\exp\{it^TX\}\exp\{it^TY\}\right] = \mathcal{X}(t;X)\mathcal{X}(t;Y)$: the characteristic function of a sum of independent random vectors factors into the product of characteristic functions.

(v) For two probability distributions P and Q, $P=Q$ if and only if $\mathcal{X}(t;P) = \mathcal{X}(t;Q)$ for all t.

Property (iv) is attractive from a NGCA standpoint, as the characteristic function factors along independent components. Suppose a random vector $X \sim P$ has independent components $X = (X_1, X_2)^T$ where $X_1 \sim P_1$ is independent of $X_2 \sim P_2$. Partition $t = (t_1, t_2)^T$ according to the dimensions of X_1 and X_2. Since $X = (X_1, 0)^T + (0, X_2)^T$, $\mathcal{X}(t;P) = \mathcal{X}(t_1;P_1)\mathcal{X}(t_2,P_2)$.

Let $X \sim P$ have a NGCA decomposition as in Definition 2.2.1. Let μ and Σ denote the mean and covariance matrix of X. Then:

$$\mathcal{X}(\Gamma s + \eta t; P) = \mathcal{X}(\Gamma s; P)\mathcal{X}(\eta t; P).$$

Since $\eta^T X \sim \mathcal{N}(\eta^T\mu, \eta^T\Sigma\eta)$, its characteristic function is well-known. We obtain:

$$\mathcal{X}(\Gamma s + \eta t; P) = \exp\left(it^T\eta^T\mu - \frac{1}{2}t^T\eta^T\Sigma\eta t\right)\mathcal{X}(\Gamma s; P). \tag{2.1}$$

Thus, projecting the data points into the non-Gaussian and Gaussian subspaces factors the characteristic function into the very specific form in (2.1). By property (v) of characteristic functions, the converse holds: if some choice of Γ and η produce the factorization of the characteristic function in (2.1) then Γ and η are projections onto independent non-Gaussian and Gaussian subspaces.

This suggests the following procedure: given Γ and η, use the following criterion to check for a NGCA decomposition:

$$\rho(\Gamma, \eta, P) = \int \left|\mathcal{X}(\Gamma s + \eta t; P) - \exp\left(it^T\eta^T\mu + \frac{1}{2}t^T\eta^T\Sigma\eta t\right)\mathcal{X}(\Gamma s; P)\right|^2 \mathrm{d}F(s,t),$$

where F is a finite-valued measure, such as a probability distribution, on $\mathbb{R}^p$ to make the integral converge. If P is a NGCA model and Γ and η span the non-Gaussian and Gaussian subspaces, then this criterion should be 0. However, we would like the converse to be true: that way we can check the criterion for different choices of Γ and η. Thus we require P to be an identifiable NGCA model (see Definition 1.2.11). Furthermore, there is no a priori relationship assumed between the non-Gaussian and Gaussian subspaces. Therefore checking this criterion amounts to a search over a d-dimensional linear subspace and a $p-d$ dimensional linear subspace, independently. By pre-whitening we obviate this difficulty and ensure the non-Gaussian and Gaussian spaces must be orthogonal (See Proposition 1.2.14 in Section 1.2).

2.2.3 The CHFNGCA estimator

Given $\Gamma \in \mathbb{R}^{p\times d}$ and $\eta \in \mathbb{R}^{p\times(p-d)}$ a naive criterion for detecting non-Gaussian and Gaussian components is:

$$\rho(\Gamma, \eta, P) = \int \left| \mathcal{X}(\Gamma s + \eta t; P) - \exp\left(it^T\eta^T\mu + \frac{1}{2}t^T\eta^T\Sigma\eta t \right) \mathcal{X}(\Gamma s; P) \right|^2 \mathrm{d}F(s,t).$$

If P is a NGCA model with a known positive definite covariance matrix, then by Proposition 1.2.14 we can without loss of generality assume that the mean of P is 0 and the covariance is I_p via the whitening transformation; we can also make the assumption that the Gaussian and non-Gaussian subspaces are orthogonal. Let $\Gamma \in \mathbb{R}^{p\times d}$ be orthogonal and let $\Gamma_\perp \in \mathbb{R}^{p\times(p-d)}$ be an orthogonal matrix satisfying $\Gamma^T\Gamma_\perp = 0$. Choosing the standard Gaussian density on $\mathbb{R}^p$ yields the following criterion for checking whether Γ spans the non-Gaussian space:

$$\rho(\Gamma, P) = \iint \left| \mathcal{X}(\Gamma s + \Gamma_\perp t; P) - \exp(-\|t\|_2^2/2)\, \mathcal{X}(\Gamma s; P) \right|^2 \phi_d(s)\phi_{p-d}(t)\,\mathrm{d}s\,\mathrm{d}t. \tag{2.2}$$

Here ϕ_d and ϕ_{p-d} are the standard Gaussian probability density functions in dimension d and $p-d$ respectively. Though it is not necessary to use the standard Gaussian density as a weighting function, this choice is convenient: we can write down an alternate form for $\rho(\Gamma, P)$ as an integral over P in closed form that only depends on Γ (see Proposition 2.3.3). Furthermore, the rotational invariance of the Gaussian distribution means that the criterion can be viewed as a function of Γ on the d-dimensional *Grassmann manifold* $\mathfrak{G}_{d,p}$, the set of all d-dimensional linear subspaces of $\mathbb{R}^p$. This is the natural parameter space for NGCA.

Let Γ be $p\times d$ and orthogonal. A point on $\mathfrak{G}_{d,p}$ is a d-dimensional linear subspace; it can be represented by the orbit $\{\Gamma U\}$ as U ranges over $d\times d$ orthogonal matrices. We demonstrate that $\rho(\Gamma, P)$ is constant on this orbit, making it a function on the Grassmann manifold. Moreover, if P has a first moment, $\rho(\Gamma, P)$ is a continuous function on the Grassmann manifold with respect to the arc length metric (see [17], p. 337 for the definition of this metric).

Proposition 2.2.2. *For $\rho(\Gamma, P)$ as defined in (2.2):*

(i) $\rho(\Gamma, P)$ *is bounded.*

(ii) *For any* $d \times d$ *orthogonal matrix* U, $\rho(\Gamma U, P) = \rho(\Gamma, P)$. *Therefore* $\rho(\Gamma, P)$ *is a function on the Grassmann manifold* $\mathfrak{G}_{d,p}$.

(iii) $\rho(\Gamma, P)$ *is a continuous function on the Grassmann manifold with respect to the arc length metric* ν.

Proof. (i) Since $\left|\mathcal{X}(s; P)\right| \leq 1$ for all s and P, we have $\left|\rho(\Gamma, P)\right| \leq 2$.

(ii) Calculate:

$$\rho(\Gamma U, P) = \iint \left|\mathcal{X}(\Gamma U s + \Gamma_\perp t; P) - \exp(-\|t\|_2^2/2)\, \mathcal{X}(\Gamma U s; P)\right|^2 \phi_d(s)\phi_{p-d}(t)\, \mathrm{d}s\, \mathrm{d}t.$$

Change variables to $r = Us$. The determinant of U in absolute value is 1, therefore,

$$\rho(\Gamma U, P) = \iint \left|\mathcal{X}(\Gamma r + \Gamma_\perp t; P) - \exp(-\|t\|_2^2/2)\, \mathcal{X}(\Gamma r; P)\right|^2 \phi_d(U^T r)\phi_{p-d}(t)\, \mathrm{d}s\, \mathrm{d}t.$$

But $\phi_d(U^T r) = \phi_d(r)$ by the rotational invariance of the Gaussian distribution. Hence $\rho(\Gamma U, P) = \rho(\Gamma, P)$. This same argument can be used to show that $\rho(\Gamma, P)$ is invariant to an orthogonal change of basis of $\Gamma_\perp$.

(iii) Let Γ and Γ_1 be two $p \times d$ orthogonal matrices, and let $\Gamma_\perp$ and $\Gamma_{1_\perp}$ be $p \times (p-d)$ orthogonal matrices which satisfy $\Gamma^T \Gamma_\perp = 0$, $\Gamma_1^T \Gamma_{1_\perp} = 0$. We can find $B \in \mathbb{R}^{(p-d)\times d}$ such that

$$\begin{pmatrix}\overline{\Gamma}_1 & \overline{\Gamma}_{1_\perp}\end{pmatrix} = \begin{pmatrix}\Gamma & \Gamma_\perp\end{pmatrix} \exp\left(\begin{bmatrix} 0 & -B^T \\ B & 0 \end{bmatrix}\right), \tag{2.3}$$

where $\operatorname{span}(\overline{\Gamma}_1) = \operatorname{span}(\Gamma_1)$ and $\operatorname{span}(\overline{\Gamma}_{1_\perp}) = \operatorname{span}(\Gamma_{1_\perp})$; furthermore $\nu(\Gamma, \Gamma_1) = \|B\|_F$ (the arc length metric ν is technically defined on the Grassmann manifold $\mathfrak{G}_{d,p}$, but we take Γ and Γ_1 to be representatives of their respective column spaces). The details of this construction can be found in [21]. Since $\rho(\overline{\Gamma}_1, P) = \rho(\Gamma_1, P)$, by Proposition B.3.1 on page 89 of Appendix B we have the inequality:

$$\begin{aligned}\left|\rho(\Gamma, P) - \rho(\Gamma_1, P)\right| \leq 4 \int \phi_d(s)\phi_{p-d}(t) \Big\{ &\left|\mathcal{X}(\Gamma s + \Gamma_\perp t; P) - \mathcal{X}(\overline{\Gamma}_1 s + \overline{\Gamma}_{1_\perp} t; P)\right| \\ &+ \left|\mathcal{X}(\Gamma s; P) - \mathcal{X}(\overline{\Gamma}_1 s; P)\right| \Big\} \mathrm{d}s\mathrm{d}t.\end{aligned}$$

Using the construction of $\overline{\Gamma}_1$, observe that as $\|B\|_F \to 0$, $\overline{\Gamma}_1 \to \Gamma$ in the Frobenius norm. Similarly, $\overline{\Gamma}_{1_\perp} \to \Gamma_\perp$. The continuity of characteristic functions implies $\mathcal{X}(\Gamma s + \Gamma_\perp t; P) \to$

$\mathcal{X}(\overline{\Gamma}_1 s + \overline{\Gamma}_{1_\perp} t; P)$ pointwise in s and t. Since the integrand is bounded by 2, then by the Dominated Convergence Theorem,

$$\int \phi_d(s)\phi_{p-d}(t)\big|\mathcal{X}(\Gamma s + \Gamma_\perp t; P) - \mathcal{X}(\overline{\Gamma}_1 s + \overline{\Gamma}_{1_\perp} t; P)\big|\mathrm{d}s\mathrm{d}t \to 0.$$

By the same arguments we also have:

$$\int \phi_d(s)\phi_{p-d}(t)\big|\mathcal{X}(\Gamma s; P) - \mathcal{X}(\overline{\Gamma}_1 s; P)\big|\mathrm{d}s\mathrm{d}t \to 0.$$

Thus continuity on $\mathfrak{G}_{d,p}$ is established.

□

The next proposition shows that $\rho(\Gamma, P)$ has desirable properties when P is an identifiable NGCA distribution with non-Gaussian subspace spanned by Γ:

Proposition 2.2.3. *Let P_0 be an identifiable NGCA model with $\mathbb{E}_P(X) = 0$, $\mathrm{Cov}_P(X) = I_p$, and d-dimensional non-Gaussian subspace spanned by the columns of the orthogonal matrix Γ_0. Then:*

(i) $\rho(\Gamma, P_0) = 0$ *if and only if* $\mathrm{span}(\Gamma) = \mathrm{span}(\Gamma_0)$.

(ii) Γ_0 *is a strong minimizer of* $\rho(\Gamma, P_0)$*: for all* $\delta > 0$,

$$\inf_{\nu(\Gamma,\Gamma_0)\geq\delta} \rho(\Gamma, P_0) > 0,$$

where ν is the arc length metric on $\mathfrak{G}_{d,p}$.

Proof. (i) If Γ orthogonal satisfies $\mathrm{span}(\Gamma) = \mathrm{span}(\Gamma_0)$, and if $\Gamma_\perp \in \mathbb{R}^{p\times(p-d)}$ orthogonal satisfies $\Gamma^T\Gamma_\perp = 0$, then Γ spans the non-Gaussian space and $\Gamma_\perp$ spans the Gaussian space. Thus:

$$\mathcal{X}(\Gamma s + \Gamma_\perp t; P_0) = \exp(-\|t\|_2^2/2)\,\mathcal{X}(\Gamma s; P_0),$$

for all (s, t) in $\mathbb{R}^p$. Therefore $\rho(\Gamma, P_0) = 0$.

Now suppose a $p \times d$ orthogonal matrix Γ satisfies $\rho(\Gamma, P_0) = 0$. Then $\mathcal{X}(\Gamma s + \Gamma_\perp t; P_0) - \exp(-\|t\|_2^2/2)\,\mathcal{X}(\Gamma s; P_0) = 0$ for all s and t except possibly a set of measure zero under the standard normal distribution. However, since the characteristic function is continuous, $\mathcal{X}(\Gamma s + \Gamma_\perp t; P_0) - \exp(-\|t\|_2^2/2)\,\mathcal{X}(\Gamma s; P_0)$ is continuous in s and t which implies the equality holds everywhere. That is, for all s and t:

$$\mathcal{X}(\Gamma s + \Gamma_\perp t; P_0) = \exp(-\|t\|_2^2/2)\,\mathcal{X}(\Gamma s; P_0).$$

For $X \sim P$, the left hand side is the characteristic function of $(\Gamma^T X, \Gamma_\perp^T X)$ while the right hand side is the characteristic function of (V, G) where $V \in \mathbb{R}^d$ is non-Gaussian and independent of $G \in \mathbb{R}^{p-d}$ which is Gaussian. Thus Γ and $\Gamma_\perp$ form a d-dimensional NGCA decomposition of X. By the assumption of identifiability, we must have $\text{span}(\Gamma) = \text{span}(\Gamma_0)$.

(ii) Suppose there exists $\delta > 0$ such that $\inf_{\nu(\Gamma,\Gamma_0)\geq\delta} \rho(\Gamma, P_0) = 0$. Then there exists a sequence of subspace parameters $\{\Gamma_j\}$ such that $\nu(\Gamma_0, \Gamma_j) \geq \delta$ and $\rho(\Gamma_j, P_0) \to 0$. Since the Grassmann manifold $\mathfrak{G}_{d,p}$ equipped with the arc length metric ν is compact, there exists a convergent subsequence $\{\Gamma_{j_k}\} \to \Gamma$. By continuity of ρ (Proposition 2.2.2) we must have $\rho(\Gamma, P_0) = 0$ which implies $\text{span}(\Gamma) = \text{span}(\Gamma_0)$. This implies $\nu(\Gamma_0, \Gamma_{j_k}) \to 0$, which contradicts $\nu(\Gamma_0, \Gamma_{j_k}) \geq \delta$. Therefore,

$$\inf_{\nu(\Gamma,\Gamma_0)\geq\delta} \rho(\Gamma, P_0) > 0.$$

□

Propositions 2.2.2 and 2.2.3 motivate the following characteristic function based method for NGCA, which we call CHFNGCA: given an i.i.d. sample $X_1, \ldots, X_n$ from an identifiable NGCA model P_0, if $\mu = \mathbb{E}(X_1)$ and $\Sigma = \text{Cov}(X_1)$ are known then we can assume, without loss of generality, that $\mu = 0$ and $\Sigma = I_p$. We estimate the whitened non-Gaussian subspace by:

$$\widehat{\Gamma}_n = \operatorname*{argmin}_{\Gamma \in \mathfrak{G}_{d,p}} \rho(\Gamma, \widehat{P}_n),$$

for $\widehat{P}_n$ the empirical distribution of the $\tilde{X}_i$. In the situation where μ and Σ are unknown (the more likely scenario), then we empirically pre-whiten the data: $\hat{X}_i = \hat{\Sigma}^{-1/2}\,(X_i - \hat{\mu})$ for consistent estimators $\hat{\mu}$ and $\hat{\Sigma}$. For example, we could set $\hat{\mu} = \sum_i X_i/n$ (the sample mean), and, for $n > p$, $\hat{\Sigma} = \sum_i (X_i - \hat{\mu})(X_i - \hat{\mu})^T/n$ the sample covariance. We then estimate the whitened non-Gaussian subspace by:

$$\widehat{\widehat{\Gamma}} = \operatorname*{argmin}_{\Gamma \in \mathfrak{G}_{d,p}} \rho(\Gamma, \widehat{P}_n(\hat{\Sigma}, \hat{\mu})),$$

where $\widehat{P}_n(\hat{\Sigma}, \hat{\mu})$ is the empirical distribution of the $\hat{X}_i$. In Section 2.3 we show that $\widehat{\Gamma}_n$ and $\widehat{\widehat{\Gamma}}_n$ are both consistent and asymptotically normal.

The optimization problem that defines $\widehat{\Gamma}_n$ or $\widehat{\widehat{\Gamma}}_n$ is not convex, due to the presence of trigonometric exponential functions in computing the empirical characteristic function. Practical algorithms for implementing CHFNGCA run the risk of getting trapped in local minima. Iterative optimization routines that mimic e.g. gradient descent on the Grassmann manifold (such as those described in [17]) may have to be started at points close to the optimum (these points could be estimates from other NGCA procedures).

2.3 Consistency and asymptotic normality

This section contains the main theorems on the statistical performance of CHFNGCA. Under i.i.d. sampling from an identifiable NGCA distribution, CHFNGCA asymptotically recovers the true non-Gaussian subspace. When the population mean and covariance are known, the estimator is consistent; furthermore, the estimator exhibits $\sqrt{n}$-asymptotic normality when the distribution possesses finite third moments (this condition is equivalent to the unknown non-Gaussian distribution possessing finite third moments). Consistency and $\sqrt{n}$-asymptotic normality continue to hold when the population mean and covariance are unknown (the latter property under finite fourth moments). The idea is to center and re-scale the data using the sample mean and sample covariance matrix, and apply the characteristic function method to the "pre-whitened" data. We show how estimation of the population mean and population covariance contribute terms to the asymptotic variance of the estimate.

2.3.1 Known population mean and covariance.

When the data are distributed according to an identifiable NGCA model with known mean and covariance matrix, we can without loss of generality assume that the mean is 0 and the covariance matrix is the identity. Our estimating criterion is $\rho(\Gamma, P)$ which was defined in (2.2). Given a sample $X_1, \ldots, X_n$ the estimate of the non-Gaussian space is:

$$\widehat{\Gamma}_n = \operatorname*{argmin}_{\Gamma \in \mathfrak{G}_{d,p}} \rho(\Gamma, \widehat{P}_n), \tag{2.4}$$

where $\mathfrak{G}_{d,p}$ is the d-dimensional Grassmann manifold in $\mathbb{R}^p$ and $\widehat{P}_n$ is the empirical distribution of the sample.

$\widehat{\Gamma}_n$ is consistent for the non-Gaussian subspace. This result is stated as Theorem 2.3.2. To prove it we need uniform control of the random function $\rho(\Gamma, \widehat{P}_n)$, stated as Lemma 2.3.1:

Lemma 2.3.1. *Let $X_1, \ldots, X_n$ be drawn i.i.d. from some distribution P. Let $\widehat{P}_n$ be the empirical distribution. Then:*

$$\sup_{\Gamma \in \mathfrak{G}_{d,p}} \left| \rho(\Gamma, P) - \rho(\Gamma, \widehat{P}_n) \right| = o_{P^*}(1),$$

where the notation P^* refers to P-outer probability (see [41], p. 6).

Note that we work in outer probability to avoid measurability issues that may arise when taking the supremum over an uncountable set. See the proof for more details.

Proof. See Appendix B, page 61. □

Theorem 2.3.2 (Consistency when μ and Σ are known.)**.** *Let $X_1, \ldots, X_n$ be i.i.d. p-dimensional random vectors with common distribution P_0, an identifiable d-dimensional NGCA model with zero mean and identity covariance. Let the d-dimensional non-Gaussian* *subspace parameter be spanned by the columns of* Γ_0*, a* $p \times d$ *orthogonal matrix. Then* $\widehat{\Gamma}$ *is consistent for* Γ_0 *with respect to the arc length metric* ν *on* $\mathfrak{G}_{d,p}$*, i.e.* $\nu(\widehat{\Gamma}, \Gamma_0) = o_{P^*}(1)$.

Proof. By Proposition 2.2.3 for all $\delta > 0$ we have:

$$\inf_{\nu(\Gamma,\Gamma_0)\geq\delta} \rho(\Gamma, P_0) > 0.$$

Let $\epsilon(\delta)$ denote the value of the infimum in the above display. If there exists Γ such that $\rho(\Gamma, P_0) < \epsilon(\delta)$ we must have $\nu(\Gamma, \Gamma_0) < \delta$. We now show this is true for $\Gamma = \widehat{\Gamma}_n$ with probability tending to 1.

Define $\Delta_n = \sup_{\Gamma\in\mathfrak{G}_{d,p}} \left|\rho(\Gamma, \widehat{P}_n) - \rho(\Gamma, P_0)\right|$. By Lemma 2.3.1, $\Delta_n \overset{P^*}{\to} 0$. Using the fact that $\widehat{\Gamma}_n$ is the minimizer of $\Gamma \in \mathfrak{G}_{d,p}\rho(\Gamma, \widehat{P}_n)$ and $\rho(\Gamma_0, P_0) = 0$ we obtain the chain of inequalities:

$$\begin{aligned}\rho\left(\widehat{\Gamma}_n, P_0\right) &\leq \rho(\widehat{\Gamma}_n, \widehat{P}_n) + \Delta_n \\ &\leq \rho(\Gamma_0, \widehat{P}_n) + \Delta_n \\ &\leq 2\Delta_n.\end{aligned}$$

This suffices to show $\nu(\widehat{\Gamma}_n, \Gamma_0) \leq \delta$ with (outer) probability tending to 1. □

Asymptotic normality.

Provided the non-Gaussian component has finite third moments, the estimate $\widehat{\Gamma}_n$ defined in (2.4) is $\sqrt{n}$-asymptotically normal when the population mean and covariance are known. Before we present the main result, it is necessary to introduce some notation.

The proof of asymptotic normality relies heavily on the Grassmannian being a smooth, differentiable manifold. To state the result we must parameterize the Grassmann manifold in a suitable way, one that allows us to take derivatives in that space. We describe the parameterization now. Let the span of the columns of the $p \times d$ orthogonal matrix Γ_0 represent a given base subspace in $\mathfrak{G}_{d,p}$. For a $(p-d) \times d$ matrix B define the mapping $\Gamma(B)$ by:

$$\Gamma(B) = (\Gamma_0\ \ \Gamma_{0_\perp}) \exp\left(\begin{bmatrix} 0 & -B \\ B & 0 \end{bmatrix}\right) J_{p,d},$$

where $\Gamma_{0_\perp}$ is a $p \times (p-d)$ orthogonal matrix which satisfies $\Gamma_{0_\perp}^T \Gamma_0 = 0$, and $J_{p,d}$ is a $p \times d$ matrix consisting of the first d columns of the p-dimensional identity matrix. Some relevant properties are:

(i) $\Gamma(0) = \Gamma_0$.

(ii) $\Gamma(B)$ is a $p \times d$ orthogonal matrix for all B; this follows from the fact that e^X is orthogonal for any square skew-symmetric matrix X. Thus any $\Gamma(B)$ can be identified as a point on the Grassmann manifold.

(iii) Given another d-dimensional subspace represented by Γ_1, there exists a $(p-d) \times d$ matrix B_1 such that $\Gamma_1 = \Gamma(B_1)$. Furthermore, for the arc length metric ν on the Grassmann manifold, B_1 satisfies $\nu(\Gamma_1, \Gamma_0) = \|B_1\|_F$ where $\|\cdot\|_F$ is the Frobenius norm (computing B_1 is an important algorithmic task; see [21] for more details). Therefore $\Gamma(B)$ maps onto the Grassmannian.

Take Γ_0 to be the d-dimensional non-Gaussian subspace of some NGCA model P_0 from which we draw n i.i.d. samples. For $\widehat{\Gamma}_n$ computed as in (2.4), by property (iii) there exists $\widehat{B}_n$ such that $\widehat{\Gamma}_n = \Gamma(\widehat{B}_n)$. Theorem 2.3.4 provides an asymptotic expansion for $\widehat{B}_n$ into a sum of i.i.d. random variables. We need one more result to write down the correct influence function: an alternative formula for the criterion $\rho(\Gamma, P)$ defined in (2.2).

Proposition 2.3.3. *$\rho(\Gamma, P) = \iint r(x, y, \Gamma) \mathrm{d}P(x) \mathrm{d}P(y)$ where:*

$$\begin{aligned} r(x,y,\Gamma) &= \exp(-\frac{1}{2}\|x-y\|_2^2) \\ &\quad + \exp\left(-\frac{1}{2}\|\Gamma^T(x-y)\|_2^2\right)\left[\left(\frac{1}{3}\right)^{\frac{p-d}{2}} - 2\left(\frac{1}{2}\right)^{\frac{p-d}{2}} \exp\left(-\frac{1}{4}\|\Gamma_\perp^T x\|_2^2\right)\right]. \end{aligned} \tag{2.5}$$

(Note: $\|\Gamma_\perp^T x\|_2^2 = \|x\|_2^2 - \|\Gamma^T x\|_2^2$. This emphasizes that ρ is indeed a function over the $p \times d$ Grassmann manifold, and we can ignore the orthogonal complement. To compute $\widehat{\Gamma}_n$ and $\widehat{\tilde{\Gamma}}_n$ we only need to optimize over a d-dimensional subspace.)

Proof. The proof is given in Appendix B. □

We now state the theorem:

Theorem 2.3.4. *Let P_0 be an identifiable NGCA model with zero mean and identity covariance. Let the $p \times d$ orthogonal matrix Γ_0 represent the non-Gaussian subspace, and let F be the non-Gaussian distribution of $\Gamma_0^T X_1$. Assume P_0 has finite third moments. Then:*

$$\widehat{B}_n = 3^{\frac{p-d}{2}} \frac{3}{2} \frac{1}{n} \sum_{i=1}^{n} \psi(X_i, \Gamma_0, P_0) M(F)^{-1} + o_P\left(n^{-\frac{1}{2}}\right),$$

where

$$\psi(x, \Gamma(B), P) = \mathbb{E}_P\left[\nabla_B r(x, X, \Gamma(B))\right] + \mathbb{E}_P\left[\nabla_B r(X, x, \Gamma(B))\right],$$

and

$$M(F) = \int \left(\nabla \mathcal{X}(s;F) + s\mathcal{X}(s;F)\right) \overline{\left(\nabla \mathcal{X}(s;F) + s\mathcal{X}(s;F)\right)}^T \phi_d(s)\mathrm{d}s;$$

here $\nabla \mathcal{X}(s;F) = \mathbb{E}_F \left(iXe^{is^TX} \right)$ *is the gradient of the characteristic function of* F *at* s*.*

Note: the identifiability condition ensures that the inverse of $M(F)$ as defined in the Theorem exists. See the proof of the theorem in Appendix B.
Asymptotic normality of $\widehat{B}_n$ is given in the following corollary:

Corollary 2.3.5. *Let* $\mathrm{vec}(\widehat{B}_n)$ *be the* $d(p-d)$*-dimensional vector obtained by stacking the columns of* $\widehat{B}_n$ *on top of each other (see e.g. [23]). Under the assumptions of Theorem 2.3.4,*

$$\sqrt{n}\mathrm{vec}(\widehat{B}_n) \xrightarrow{d} \mathcal{N}(0, C(\Gamma_0, P_0, F)),$$

where:

$$C(\Gamma_0, P_0, F) = 3^{p-d}\frac{9}{4}\left[M(F)^{-1} \otimes I_{p-d}\right] \mathrm{Cov}_{P_0}\left(\mathrm{vec}\left(\psi(X_1, \Gamma_0, P_0)\right)\right) \left[M(F)^{-1} \otimes I_{p-d}\right];$$

here $\otimes$ *denotes the Kronecker product for matrices [23].*

Proof. Using the linearity of the vec operator and the formula $\mathrm{vec}(ABC) = \left(C^T \otimes A\right)\mathrm{vec}(B)$ [23], we have from Theorem 2.3.4

$$\sqrt{n}\mathrm{vec}(\widehat{B}_n) = 3^{(p-d)/2}\frac{3}{2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[M(F)^{-1} \otimes I_{p-d}\right]\mathrm{vec}(\psi(X_i, \Gamma_0, P_0) + o_P(1).$$

Each element of $\psi(x, \Gamma(B), P)$ is up to a constant bounded by $\|x\|_2$ (Lemma B.3.2 in the appendix) which implies $\mathrm{Cov}_{P_0}\left(\mathrm{vec}\left(\psi(X_1, \Gamma_0, P_0)\right)\right)$ exists. The corollary follows from the multivariate central limit theorem. □

2.3.2 Unknown mean and covariance.

If the data are distributed according to a NGCA distribution with unknown mean and covariance, then the relationship between the Gaussian and non-Gaussian subspaces is also unknown. To deal with this, we estimate the mean and covariance and pre-whiten the data empirically. We then apply CHNGCA to the empirically pre-whitened data. We require a more sophisticated analysis to deal with the variability introduced by estimating the mean and covariance.

This is the setup: suppose $X_1, \ldots, X_n$ are a sample from an identifiable NGCA model with unknown mean μ_0 and unknown positive definite covariance matrix Σ_0. Let $\Gamma_0 \in \mathbb{R}^{p\times d}$ be an orthogonal matrix whose columns span the non-Gaussian subspace. The non-Gaussian subspace corresponding to the whitened data points $\Sigma_0^{-1/2}(X_i - \mu_0)$, $i = 1, \ldots, n$ is given by $\mathrm{span}(\Sigma_0^{1/2}\Gamma_0)$; we call this subspace the whitened non-Gaussian subspace. Our goal is to estimate it.

Let $\hat{\mu}$ and $\hat{\Sigma}$ be consistent estimators of μ_0 and Σ_0. For example, let $\hat{\mu} = \sum_i X_i/n$ (the sample mean) and $\hat{\Sigma} = \sum(X_i - \hat{\mu})(X_i - \hat{\mu})^T/n$ (the sample covariance). For $p < n$, if $\hat{\Sigma}^{-1}$ exists, we empirically pre-whiten the data via the following transformation: $\hat{X}_i = \hat{\Sigma}^{-1/2}(X_i - \hat{\mu}), i = 1, \ldots, n$. Then using the criterion $\rho(\Gamma, P)$ defined in (2.2), we form an estimate of the whitened non-Gaussian subspace by:

$$\widehat{\tilde{\Gamma}} = \underset{\Gamma \in \mathfrak{G}_{d,p}}{\operatorname{argmin}}\ \rho(\Gamma, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})), \tag{2.6}$$

where $\widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})$ is the empirical distribution of $\hat{X}_1, \ldots, \hat{X}_n$.

We will state a theorem which shows $\widehat{\tilde{\Gamma}}$ is consistent for the whitened non-Gaussian subspace. To prove it, we need uniform boundedness of ρ when the population mean and covariance are estimated:

Lemma 2.3.6. *Let $X_1, \ldots, X_n$ be i.i.d. p-dimensional random vectors distributed according to P, with mean μ and positive definite covariance Σ. Let $\widehat{\mu}$ and $\widehat{\Sigma}$ be consistent estimators of μ, Σ respectively. Define $\tilde{X}_i = \Sigma^{-\frac{1}{2}}(X_i - \mu)$ and $\hat{X}_i = \widehat{\Sigma}^{-\frac{1}{2}}(X_i - \widehat{\mu})$ for $i = 1, \ldots, n$. Denote by $\widehat{\tilde{P}}_n(\Sigma, \mu)$ the empirical distribution of $\tilde{X}_1, \ldots, \tilde{X}_n$ and by $\widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})$ the empirical distribution of $\hat{X}_1, \ldots, \hat{X}_n$.Then:*

$$\sup_{\Gamma \in \mathfrak{G}_{d,p}} \left|\rho(\Gamma, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})) - \rho(\Gamma, \widehat{\tilde{P}}_n(\Sigma, \mu))\right| = o_{P^*}(1).$$

Proof. See the appendix. □

We now state the theorem.

Theorem 2.3.7. *Let $X_1, \ldots, X_n$ be i.i.d.p-dimensional random vectors with common distribution P_0, an identifiable d-dimensional NGCA model with unknown mean μ_0 and unknown positive definite covariance matrix Σ_0. Let $\tilde{P}_0$ be the distribution of the whitened data $\tilde{X}_i = \Sigma_0^{-1/2}(X_i - \mu_0)$, and let the columns of the $p \times d$ orthogonal matrix $\tilde{\Gamma}_0$ span the whitened non-Gaussian subspace. Then for $\widehat{\tilde{\Gamma}}_n$ defined in (2.6), we have $\nu(\widehat{\tilde{\Gamma}}_n, \tilde{\Gamma}_0) = o_{P^*}(1)$, where ν is the arc length metric on the Grassmann manifold $\mathfrak{G}_{d,p}$.*

Proof. Let $\widehat{\tilde{P}}_n(\Sigma_0, \mu_0)$ be the empirical distribution on $\tilde{X}_1, \ldots, \tilde{X}_n$, and consider the unobservable estimate of the whitened non-Gaussian subspace given by:

$$\widehat{\tilde{\Gamma}}_n^* = \underset{\Gamma \in \mathfrak{G}_{d,p}}{\operatorname{argmin}}\ \rho(\Gamma, \widehat{\tilde{P}}_n(\Sigma_0, \mu_0)).$$

Define the uniform bounds $\Delta_n = \sup_{\Gamma \in \mathfrak{G}_{d,p}} \left|\rho(\Gamma, \widehat{\tilde{P}}_n(\Sigma_0, \mu_0)) - \rho(\Gamma, \tilde{P}_0)\right|$ and $\Delta_n(\widehat{\Sigma}, \widehat{\mu}) = \sup_{\Gamma \in \mathfrak{G}_{d,p}} \left|\rho(\Gamma, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})) - \rho(\Gamma, \widehat{\tilde{P}}_n(\Sigma_0, \mu_0))\right|$. Recall $\rho(\tilde{\Gamma}_0, \tilde{P}_0) = 0$, $\widehat{\tilde{\Gamma}}_n^*$ is the minimizer of $\rho(\Gamma, \widehat{\tilde{P}}_n(\Sigma_0, \mu_0))$, and $\widehat{\tilde{\Gamma}}_n$ is the minimizer of $\rho(\Gamma, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu}))$. Then:

$$\begin{aligned}
\rho\left(\widehat{\tilde{\Gamma}}_n, \tilde{P}_0\right) &\leq \rho\left(\widehat{\tilde{\Gamma}}_n, \widehat{\tilde{P}}_n(\Sigma_0, \mu_0)\right) + \Delta_n \\
&\leq \rho\left(\widehat{\tilde{\Gamma}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})\right) + \Delta_n(\widehat{\Sigma}, \widehat{\mu}) + \Delta_n \\
&\leq \rho\left(\widehat{\tilde{\Gamma}}_n^*, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})\right) + \Delta_n(\widehat{\Sigma}, \widehat{\mu}) + \Delta_n \\
&\leq \rho\left(\widehat{\tilde{\Gamma}}_n^*, \widehat{\tilde{P}}_n(\Sigma_0, \mu_0)\right) + 2\Delta_n(\widehat{\Sigma}, \widehat{\mu}) + \Delta_n \\
&\leq \rho\left(\widehat{\tilde{\Gamma}}_n^*, \tilde{P}_0\right) + 2\Delta_n(\widehat{\Sigma}, \widehat{\mu}) + 2\Delta_n.
\end{aligned}$$

Δ_n and $\Delta_n(\widehat{\Sigma}, \widehat{\mu})$ tend to 0 in probability by Lemmas 2.3.1 and 2.3.6; $\rho(\widehat{\tilde{\Gamma}}^*, \tilde{P}_0)$ tends to zero in probability as we showed in the proof of Theorem (2.3.2). We conclude $\rho\left(\widehat{\tilde{\Gamma}}_n, \tilde{P}_0\right) \to 0$ in probability, and hence $\widehat{\tilde{\Gamma}}_n$ is consistent for $\tilde{\Gamma}_0$ (by the reasoning described in the proof of Theorem 2.3.2). □

Asymptotic normality.

To state the theorem for asymptotic normality under empirical pre-whitening, we first define a path on the Grassmann manifold $\mathfrak{G}_{p,d}$ starting at the whitened non-Gaussian subspace $\tilde{\Gamma}_0$ by:

$$\tilde{\Gamma}(B) = \begin{pmatrix}\tilde{\Gamma}_0 & \tilde{\Gamma}_{0_\perp}\end{pmatrix} \exp\left(\begin{bmatrix} 0 & -B \\ B & 0 \end{bmatrix}\right) J_{p,d},$$

where $\tilde{\Gamma}_{0_\perp}^T \tilde{\Gamma}_0 = 0$. Consider $\widehat{\tilde{\Gamma}}_n$ defined in (2.6) with $\hat{\mu} = \sum_i X_i/n$ (the sample mean) and $\hat{\Sigma} = \sum_i (X_i - \hat{\mu})(X_i - \hat{\mu})^T/n$ (the sample covariance). Let $\widehat{\tilde{B}}_n$ satisfy $\tilde{\Gamma}(\widehat{\tilde{B}}_n) = \widehat{\tilde{\Gamma}}_n$ with $\nu(\tilde{\Gamma}_0, \widehat{\tilde{\Gamma}}_n) = \|\widehat{\tilde{B}}_n\|_F$. The next theorem provides an asymptotic expansion for $\widehat{\tilde{B}}_n$:

Theorem 2.3.8. *Let $X_1, \ldots, X_n$ be i.i.d. p-dimensional random vectors with common distribution P_0, an identifiable d-dimensional NGCA model with unknown mean μ_0, unknown positive definite covariance matrix Σ_0, and d-dimensional non-Gaussian subspace parameter represented by Γ_0, a $p \times d$ orthogonal matrix. Assume that P_0 has finite fourth moments. Let $\tilde{\Gamma}_0$ be a $n \times p$ orthogonal matrix whose column space is equal to* $\operatorname{span}(\Sigma_0^{\frac{1}{2}} \Gamma_0)$. *Let $\tilde{X}_i = \Sigma_0^{-\frac{1}{2}}(X_i - \mu_0)$, $i = 1, \ldots, n$. Then:*

$$\widehat{\widehat{B}}_n = 3^{(p-d)/2}\frac{3}{2}\frac{1}{n}\sum_{i=1}^{n}\tilde{\psi}\left(\tilde{X}_i, \tilde{\Gamma}_0, \tilde{P}_0\middle|\mu_0, \Sigma_0\right) M(\tilde{F})^{-1} + o_P(n^{-\frac{1}{2}}),$$

where $\tilde{P}_0$ is the distribution of $\tilde{X}_1$, $\tilde{F}$ is the non-Gaussian distribution of $\tilde{\Gamma}_0^T\tilde{X}_1$, and $\tilde{\psi} \in \mathbb{R}^{(p-d)\times d}$ is given in vector form by:

$$\begin{aligned}\mathrm{vec}\left(\tilde{\psi}(x, \tilde{\Gamma}(B), P\middle|\mu, \Sigma)\right) &= \mathrm{vec}\left(\psi(x, \tilde{\Gamma}(B), P)\right) - \mathbb{E}_P\left[\nabla_x^T \mathrm{vec}\left(\psi(X, \tilde{\Gamma}(B), P)\right)\right] x \\ &\quad - \frac{1}{2}\mathbb{E}_P\left[X^T \otimes \nabla_x^T \mathrm{vec}\left(\psi(X, \tilde{\Gamma}(B), P)\right)\right]\left(\Sigma^{\frac{1}{4}} \otimes \Sigma^{-\frac{1}{4}}\right)\mathrm{vec}\left(xx^T - I_p\right).\end{aligned}$$

Proof. See Appendix B. □

The assumptions of Theorem 2.3.8 guarantee that certain quantities defined therein exist. Since we assume P_0 is identifiable, $\tilde{P}_0$ is identifiable, with 0 mean, identity covariance, subspace parameter $\tilde{\Gamma}_0$ and non-Gaussian distribution $\tilde{F}$. Thus $M(\tilde{F})$ is invertible. Since we assume P_0 has finite fourth moments, the existence of the matrices $\mathbb{E}_{\tilde{P}_0}\left[\nabla_x^T \mathrm{vec}\left(\psi(X^*, \tilde{\Gamma}_0, \tilde{P}_0)\right)\right]$ and $\mathbb{E}_{\tilde{P}_0}\left[(\tilde{X})^T \otimes \nabla_x^T \mathrm{vec}\left(\psi(\tilde{X}, \tilde{\Gamma}_0, \tilde{P}_0)\right)\right]$ is guaranteed by Lemma B.3.3.

The term $\frac{1}{n}\sum_{i=1}^n \tilde{\psi}\left(\tilde{X}_i, \tilde{\Gamma}_0, \tilde{P}_0\middle|\mu_0, \Sigma_0\right)$ is decomposed into three parts. The term $\frac{1}{n}\sum_{i=1}^n \psi(\tilde{X}_i, \tilde{\Gamma}_0, \tilde{P}_0)$ is the influence function for estimating the non-Gaussian space when μ_0 and Σ_0 are known. Estimating μ_0 by $\widehat{\mu}$ contributes the term $\mathbb{E}_{\tilde{P}_0}\left[\nabla_x^T \mathrm{vec}\left(\psi(\tilde{X}, \tilde{\Gamma}_0, \tilde{P}_0)\right)\right]\left(\frac{1}{n}\sum_{i=1}^n \tilde{X}_i\right)$; note that it does not depend on the particular value of μ_0 since each $\tilde{X}_i$ has mean 0. On the other hand, the term

$$-\frac{1}{2}\mathbb{E}_{\tilde{P}_0}\left[(\tilde{X})^T \otimes \nabla_x^T \mathrm{vec}\left(\psi(\tilde{X}, \tilde{\Gamma}_0, \tilde{P}_0)\right)\right]\left(\Sigma^{\frac{1}{4}} \otimes \Sigma^{-\frac{1}{4}}\right)\mathrm{vec}\left(\frac{1}{n}\sum_{i=1}^n \tilde{X}_i\tilde{X}_i^T - I_p\right),$$

which corresponds to estimating Σ_0 by $\widehat{\Sigma}$, does depend on the value of Σ_0.

A natural question is whether these additional terms in the influence function of $\widehat{\widehat{B}}_n$, which arise from estimating unknown nuisance parameters, cause the estimator to have larger fluctuations than if μ_0 and Σ_0 were known. We conjecture that this is the case. However, we have not thus far been able to prove that the asymptotic covariance matrix of $\widehat{\widehat{B}}_n$ is larger than that of $\widehat{B}_n$ (in the semidefinite ordering on symmetric matrices). In future work, the conjecture could be investigated through simulation studies and further theoretical analysis.

2.4 Conclusion, Future Work, Open Problems.

We have proposed and analyzed a novel method called CHFNGCA for extracting the non-Gaussian components in a NGCA setting based on characteristic functions. We provide theoretical guarantees for the performance of the estimator in an asymptotic setting, including $\sqrt{n}$-consistency, and give precise fluctuation behavior for the estimate at the $\sqrt{n}$ scale. We are not aware of any other work in the field which derives the asymptotic distribution of a NGCA method.

There is much work to be done in this area. A thorough simulation study could help gauge the practical efficacy of the method and determine the usefulness of the theory we present here. We could also compare the performance of CHFNGCA to other NGCA methods, varying the kinds of departures from normality that are present in the data, or the data's dimensionality. And there are still many open problems in NGCA itself, such as: is it possible to find consistent estimates in the high-dimensional case, when the ambient dimension p of the data is allowed to go to infinity? How can we estimate the dimension of the non-Gaussian space? And can we complete the low-dimensional theory of NGCA by producing efficient estimators of the space? We now dive into these problems in detail.

2.4.1 High-dimension.

NGCA is advertised as a method for reducing data dimensionality. However, most of the theoretical work has focused on the classical case where n, the number of samples, is much larger than p, the ambient dimension. Effectively this is a low-dimensional case. Theoretical work under asymptotic approximations where $n, p \to \infty$ (perhaps with d, the non-Gaussian dimension, bounded or growing slowly) could be more useful for practical dimensionality reduction on large datasets. Simulation results have shown the SNGCA-SDP algorithm [16] performs reasonably well in moderate dimensions, so there is some evidence that NGCA can be adapted for high-dimensional situations. However, semi-definite programs tend not to scale well in high-dimensions.

One obstacle to NGCA methods attaining good high-dimensional performance is that many require the data to be empirically pre-whitened, implicitly or explicitly assuming that the sample covariance estimator is consistent for the population covariance. However, if p/n is not small, the sample covariance matrix is not consistent in spectral norm. Therefore, the performance of these estimates could degrade significantly in high-dimension. One possible solution is the use of modern regularized approaches to covariance estimation (see [2] for just one of many examples). We might hope for instance that the covariance of the Gaussian components of the data are, in some suitable sense, low-dimensional, and we could propose regularizations to capture this low-dimensional structure.

Regularization methods have become increasingly popular in statistics for dealing with ill-posed problems, such as the high-dimensional setting. When a statistical algorithm can be posed as an optimization problem constraints can be imposed on the optimization to regularize the solution and capture desirable low-dimensional structure. A well-known successful example is the LASSO, which uses ℓ_1 penalization in linear regression to produce coefficient estimates which are sparse [39]. If suitable low-dimensional behavior could be identified in NGCA – perhaps some kind of notion of sparsity — we could modify the CHFNGCA algorithm, which is an optimization problem, to accommodate constraints that encourages solutions of a desirable form. This could move NGCA methods into more relevant high-dimensional settings.

2.4.2 Estimating the non-Gaussian dimension.

Most NGCA methods assume the dimension of the non-Gaussian space is known. However, this is seldom the case in practice. To make things worse, the identifiability of the non-Gaussian space depends on the Gaussian component being of maximal dimension (recall Theorem 1.2.12). Therefore, an accurate estimate of the true dimension of the non-Gaussian space is required for any estimate to be reliable. However, there has been very little work on this problem. A method for estimating the non-Gaussian space without knowing the dimension a priori was mentioned in [16], and Theorem 2 of that paper contains promising theoretical results for recovering the non-Gaussian space at the $\sqrt{n}$-rate. However, the method depends on accurately estimating the smallest eigenvalue of the population covariance, which is challenging in high-dimensions since the usual sample covariance matrix is inconsistent.

A dimension estimator is proposed in [37] for use with the NGCA algorithm outlined in [5]. For this algorithm, PCA is run on a set of vectors which lie close to the non-Gaussian space, and the non-Gaussian space is determined from the eigenvectors corresponding to the largest eigenvalues. The idea of the dimension estimator for a given problem is to run this NGCA algorithm on simulated pure Gaussian data of the same size and dimension a number of times to generate a histogram of the eigenvalues. Then an appropriate cut-off is chosen, e.g. the 95th percentile of the eigenvalue distribution. NGCA is performed on the data of interest, and eigenvalues below the cut-off are declared to be in the Gaussian space. The estimator is accurate on the same simulated data sets used in [5]. However, theoretical insights into this problem might aid the development of more general procedures for use in other algorithms.

2.4.3 Semiparameteric efficiency.

The NGCA model is a semiparametric model with a well-defined likelihood. An analysis of the NGCA model along the lines of [3] could yield lower bounds for estimator performance and lead to the proposal of efficient estimates. This would lead to a complete theory for NGCA in low-dimension.

The main challenge for obtaining efficient estimators is computing the efficient influence function in the model. This would likely entail consistent estimation of the score functions of the non-Gaussian components in the model. We conjecture a method based on a method for obtaining efficient estimators of the un-mixing matrix in the ICA model that was analyzed in [8]. Beginning with a $\sqrt{n}$-consistent estimate of the non-Gaussian subspace, project the data into the non-Gaussian space. Using the projected data, compute estimates of the score function of the non-Gaussian components using smoothed nonparametric function estimation techniques (such as smoothing splines). Then plug the estimates of the score functions into the likelihood of the NGCA model, and maximize the likelihood over the non-Gaussian subspace parameter. Iterate this back and forth procedure to achieve efficiency. The characteristic function estimate CHFNGCA could be used as a $\sqrt{n}$-consistent starting point for the algorithm.

Bibliography

[1] Kofi P. Adragni and R. Dennis Cook. “Sufficient dimension reduction and prediction in regression”. In: *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* 367.1906 (2009), pp. 4385–4405.

[2] Peter J. Bickel and Elizaveta Levina. “Regularized estimation of large covariance matrices”. In: *Ann. Statist.* 36.1 (2008), pp. 199–227.

[3] Peter J. Bickel et al. *Efficient and adaptive estimation for semiparametric models.* Reprint of the 1993 original. Springer-Verlag, New York, 1998.

[4] Patrick Billingsley. *Probability and measure.* Second ed. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons Inc., 1986.

[5] Gilles Blanchard et al. “In search of non-Gaussian components of a high-dimensional distribution”. In: *J. Mach. Learn. Res.* 7 (2006), pp. 247–282.

[6] G. Casella and R.L. Berger. *Statistical inference.* Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002.

[7] Aiyou Chen and Peter J. Bickel. “Consistent independent component analysis and prewhitening”. In: *IEEE Trans. Signal Process.* 53.10, part 1 (2005), pp. 3625–3632.

[8] Aiyou Chen and Peter J. Bickel. “Efficient independent component analysis”. In: *Ann. Statist.* 34.6 (2006), pp. 2825–2855.

[9] Pierre Comon. “Independent Component Analysis, a New Concept?” In: *Signal Process.* 36.3 (Apr. 1994), pp. 287–314.

[10] R. Dennis Cook. “Fisher Lecture: Dimension Reduction in Regression”. In: *Statistical Science* 22.1 (Feb. 2007), pp. 1–26.

[11] R. Dennis Cook. *Regression graphics.* Wiley Series in Probability and Statistics: Probability and Statistics. Ideas for studying regressions through graphics, A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1998.

[12] R. Dennis Cook and Liliana Forzani. “Principal Fitted Components for Dimension Reduction in Regression”. In: *Statistical Science* 23.4 (Nov. 2008), pp. 485–501.

[13] R. Dennis Cook and Sanford Weisberg. “Sliced Inverse Regression for Dimension Reduction: Comment”. In: *Journal of the American Statistical Association* 86.414 (1991), pp. 328–332.

[14] Sándor Csörgő. “Multivariate empirical characteristic functions”. In: *Z. Wahrsch. Verw. Gebiete* 55.2 (1981), pp. 203–229.

[15] Elmar Diederichs et al. “Sparse non-Gaussian component analysis.” In: *IEEE Transactions on Information Theory* 56.6 (2010), pp. 3033–3047.

[16] Elmar Diederichs et al. “Sparse non Gaussian component analysis by semidefinite programming.” In: *Machine Learning* 91.2 (2013), pp. 211–238.

[17] Alan Edelman, Tomás A. Arias, and Steven T. Smith. “The geometry of algorithms with orthogonality constraints”. In: *SIAM J. Matrix Anal. Appl.* 20.2 (1999), pp. 303–353.

[18] J. Eriksson, A. Kankainen, and V. Koivunen. “Novel Characteristic Function Based Criteria for ICA”. In: *In Proc. of the 3rd Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA2001.* 2001, pp. 108–113.

[19] J. Fox. *Applied Regression Analysis and Generalized Linear Models.* SAGE Publications, 2008.

[20] J.H. Friedman and J.W. Tukey. “A Projection Pursuit Algorithm for Exploratory Data Analysis”. In: *Computers, IEEE Transactions on* C-23.9 (1974), pp. 881–890.

[21] Kyle Gallivan et al. “Efficient Algorithms For Inferences On Grassmann Manifolds”. In: *Proceedings of the 12th IEEE Workshop on Statistical Signal Processing.* 2003, pp. 315–318.

[22] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning.* Second. Springer Series in Statistics. Data mining, inference, and prediction. Springer, New York, 2009.

[23] Harold V. Henderson and S. R. Searle. “Vec and *vech* operators for matrices, with some uses in Jacobians and multivariate statistics”. In: *Canad. J. Statist.* 7.1 (1979).

[24] Wassily Hoefdding. “The strong law of large numbers for U-statistics”. In: *Univ. of North Carolina Mimeograph Series* 302 (1961).

[25] H. Hotelling. *Analysis of a complex of statistical variables into principal components.* 1933.

[26] Peter J. Huber. “Projection Pursuit”. In: *The Annals of Statistics* 13.2 (June 1985), pp. 435–475.

[27] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* Adaptive and Learning Systems for Signal Processing, Communications and Control Series. Wiley, 2004.

[28] Aapo Hyvrinen. “Fast and robust fixed-point algorithms for independent component analysis.” In: *IEEE Transactions on Neural Networks* 10.3 (1999), pp. 626–634.

[29] “Joint low-rank approximation for extracting non-Gaussian subspaces”. In: *Signal Processing* 87.8 (2007). Independent Component Analysis and Blind Source Separation, pp. 1890 –1903.

[30] Motoaki Kawanabe. “Linear Dimension Reduction Based on the Fourth-Order Cumulant Tensor.” In: *ICANN (2)*. Vol. 3697. Lecture Notes in Computer Science. Springer, Sept. 5, 2005, pp. 151–156. URL: `http://dblp.uni-trier.de/db/conf/icann/icann2005-2.html#Kawanabe05`.

[31] Motoaki Kawanabe and Fabian J. Theis. “Estimating Non-gaussian Subspaces by Characteristic Functions”. In: *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Signal Separation.* ICA’06. Charleston, SC: Springer-Verlag, 2006, pp. 157–164.

[32] Motoaki Kawanabe et al. “A new algorithm of non-Gaussian component analysis with radial kernel functions”. English. In: *Annals of the Institute of Statistical Mathematics* 59.1 (2007), pp. 57–75.

[33] E. L. Lehmann and George Casella. *Theory of point estimation.* Second. Springer Texts in Statistics. Springer-Verlag, New York, 1998.

[34] Ker-Chau Li. “Sliced inverse regression for dimension reduction”. In: *J. Amer. Statist. Assoc.* 86.414 (1991). With discussion and a rejoinder by the author, pp. 316–342.

[35] Vladimir Panov. *Non-gaussian component analysis: New ideas, new proofs, new applications.* eng. SFB 649 discussion paper 2010,026. Berlin, 2010.

[36] Karl Pearson. “On Lines and Planes of Closest Fit to Systems of Points in Space”. In: 6.2 (1901), pp. 559–572.

[37] Fabian J. Theis, Motoaki Kawanabe, and Klaus-Robert Müller. “Uniqueness of non-Gaussianity-based dimension reduction”. In: *IEEE Trans. Signal Process.* 59.9 (2011), pp. 4478–4482.

[38] FabianJ. Theis and Motoaki Kawanabe. “Uniqueness of Non-Gaussian Subspace Analysis”. In: *Independent Component Analysis and Blind Signal Separation.* Vol. 3889. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2006, pp. 917–925.

[39] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *J. Roy. Statist. Soc. Ser. B* 58.1 (1996), pp. 267–288.

[40] A. W. van der Vaart. *Asymptotic statistics.* Vol. 3. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1998.

[41] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes.* Springer series in statistics. Springer, 1996.

Appendix A

Appendix for Chapter 1

A.1 Proofs from Section 1.2

Proof of Proposition 1.2.1. Note that without loss of generality we can assume the Gaussian component G in (1.1.1) has independent coordinates. To see this, suppose we first make no assumptions about the covariance $\mathrm{Cov}(G)$. Let $\mathrm{Cov}(G) = UDU^T$ be the eigenvalue decomposition, so that U is orthogonal and D is diagonal. Consider the NGCA decomposition of X defined by:

$$\begin{bmatrix} I_d & 0 \\ 0 & U^T \end{bmatrix} \begin{bmatrix} \Gamma^T X \\ \eta^T X \end{bmatrix} = \begin{bmatrix} V \\ U^T G \end{bmatrix}.$$

Note that $U^T G$ is still Gaussian, and $\mathrm{Cov}(U^T G) = D$, which implies $U^T G$ has independent coordinates. Since $[\Gamma \quad \eta]$ is invertible if and only if the matrix product

$$\begin{bmatrix} I_d & 0 \\ 0 & U^T \end{bmatrix} \begin{bmatrix} \Gamma^T \\ \eta^T \end{bmatrix}$$

is invertible, we can just assume that G has independent coordinates.

Let Γ_j, $j = 1, \ldots, d$ be the columns of Γ and η_k, $k = 1, \ldots, p-d$ the columns of η. Then $[\Gamma \quad \eta]$ not invertible implies that there exist $a_1, \ldots, a_d, b_1, \ldots, b_{p-d}$ not all 0 such that

$$\sum_{j=1}^{d} a_j \Gamma_j + \sum_{k=1}^{p-d} b_k \eta_k = 0.$$

If all the b_k's are equal to 0, then this implies that $\dim(\mathrm{span}(\Gamma)) < d$; if all the a_j's are equal to 0 then $\dim(\mathrm{span}(\eta)) < p - d$. The third possibility is that at least one $b_k \neq 0$ and at least one $a_j \neq 0$. Then $\sum_j a_j \Gamma_j = -\sum_k b_k \eta_k$ and neither side of the equality is equal to 0, which implies $\mathrm{span}(\Gamma)$ and $\mathrm{span}(\eta)$ intersect in at least a one-dimensional subspace. We now slightly recast this equality: for it to hold, there must exist an index k', $c \in \mathbb{R}^d$ with $c \neq 0$ and $d \in \mathbb{R}^{p-d-1}$ such that

$$\eta_{k'} = \Gamma c + \eta_{-k'} d,$$

where $\eta_{-k'}$ is the $p \times (p-d-1)$ matrix consisting of all columns of η except for k'. Therefore $G_{k'} = \eta_{k'}^T X$ can be written:

$$G_{k'} = c^T V + d^T G_{-k'},$$

where $G_{-k'}$ is the random vector containing all coordinates of G except the k'th. Note that, since $c^T V = G_{k'} - d^T G_{-k'}$ and $G_{k'}$ and $d^T G_{-k'}$ are independent with finite variances, we must have $\mathrm{Var}(c^T V) < \infty$. Therefore, $\mathrm{Cov}(G_{k'}, c^T V)$ exists, which implies by independence:

$$\begin{aligned} \mathrm{Var}(G_{k'}) &= \mathrm{Cov}(G_{k'}, c^T V) + \mathrm{Cov}(G_{k'}, d^T G_{-k'}) \\ &= 0. \end{aligned}$$

Hence, $c^T V = -d^T G_{-k'}$ which implies

$$\mathrm{Var}(c^T V) = -\mathrm{Cov}(c^T V, d^T G_{-k'}) = 0.$$

We conclude that $c^T V = k$ for some k with probability 1.

□

Proof of Proposition 1.2.4. If $[\Gamma \quad \eta]$ is invertible then so is the matrix product $[\Gamma \quad \eta]\,[\Gamma \quad \eta]^T = \Gamma\Gamma^T + \eta\eta^T$. By the decomposition (1.1.1) we have

$$\left(\Gamma\Gamma^T + \eta\eta^T\right) X = \Gamma V + \eta G,$$

whence we obtain

$$X = \left(\Gamma\Gamma^T + \eta\eta^T\right)^{-1} \Gamma V + \left(\Gamma\Gamma^T + \eta\eta^T\right)^{-1} \eta G.$$

Set $V' = V$ and $G' = G$; and set $\bar{\Gamma} = \left(\Gamma\Gamma^T + \eta\eta^T\right)^{-1} \Gamma$ and $\bar{\eta} = \left(\Gamma\Gamma^T + \eta\eta^T\right)^{-1} \eta$. Since $\left[\bar{\Gamma} \quad \bar{\eta}\right] = \left([\Gamma \quad \eta]^T\right)^{-1}$, it is invertible. Therefore $\mathrm{span}(\bar{\Gamma})$ has dimension d and $\mathrm{span}(\bar{\Gamma})^\perp$ has dimension $p-d$. Notice that

$$\mathrm{span}(\bar{\Gamma})^\perp = \left(\Gamma\Gamma^T + \eta\eta^T\right) \mathrm{span}(\Gamma)^\perp = \eta\eta^T \mathrm{span}(\Gamma)^\perp.$$

Hence $\mathrm{span}(\bar{\Gamma})^\perp \subseteq \mathrm{span}(\eta)$. Since $\mathrm{span}(\eta)$ has dimension $p-d$, it follows that $\mathrm{span}(\bar{\Gamma})^\perp = \mathrm{span}(\eta)$. The subspace equality $\mathrm{span}(\bar{\eta})^\perp = \mathrm{span}(\Gamma)$ is proved the exact same way.

To prove the converse, suppose $X = \bar{\Gamma} V' + \bar{\eta} G'$ where $\left[\bar{\Gamma} \quad \bar{\eta}\right]$ is invertible. Pick η so that its columns span the same subspace as $\mathrm{span}(\bar{\Gamma})^\perp$; thus η has rank $p-d$ (this is possible since $\bar{\Gamma}$ must have rank d to meet the invertibility condition). Pick Γ so that its columns span the subspace $\mathrm{span}(\bar{\eta})^\perp$; then Γ has rank d. Since $\mathrm{span}(\bar{\Gamma})$ and $\mathrm{span}(\bar{\eta})$ do not intersect, neither do $\mathrm{span}(\Gamma)$ and $\mathrm{span}(\eta)$, which implies $[\Gamma \quad \eta]$ is invertible. Compute $[\Gamma \quad \eta]^T X$:

$$\begin{bmatrix}\Gamma^T X\\ \eta^T X\end{bmatrix} = \begin{bmatrix}\Gamma^T \bar{\Gamma} V'\\ \eta^T \bar{\eta} G'\end{bmatrix}.$$

Set $V = \Gamma^T \bar{\Gamma} V'$ and $G = \eta^T \bar{\eta} G'$. Then V is non-Gaussian and independent of Gaussian G. □

Proof of Proposition 1.2.5. This proof borrows substantially from the proof of Proposition 2 in Appendix A.2 in [5]. Under mild regularity conditions that allow for differentiation under the integral sign, we have

$$\int \nabla g(x) p(x) dx = -\int g(x) \nabla p(x) dx = -\int g(x) \nabla \log p(x)\ p(x) dx.$$

Compute $\nabla \log p(x)$:

$$\begin{aligned}\nabla \log p(x) &= \nabla \log f(\Gamma^T X) + \nabla \log \phi_{\Delta_G}(\eta^T x)\\ &= \Gamma^T \frac{\nabla f(\Gamma^T x)}{f(\Gamma^T X)} - \eta \Delta_G^{-1} \eta^T X,\end{aligned}$$

since $\log \phi_{\Delta_g}(y)$ is proportional to $-\frac{1}{2}\|\Delta_G^{-1/2} y\|_2^2$. Hence,

$$\mathbb{E}\left[\nabla g(X)\right] = -\Gamma \int \frac{\nabla f(\Gamma^T x)}{f(\Gamma^T x)} p(x) dx + \eta \Delta_G^{-1} \eta^T \mathbb{E}\left[X g(X)\right].$$

Re-arrange the equality to complete the proof. □

Proof of Proposition 1.2.7. Write $N = \Delta^{\frac{1}{2}} Z$ where $Z \sim \mathcal{N}(0, I_p)$. For any linear subspace $\mathcal{S}$ we can decompose Z by:

$$\begin{aligned}Z &= \Pi_{\mathcal{S}} Z + \Pi_{\mathcal{S}^\perp} Z\\ &= Z_1 + Z_2,\end{aligned}$$

where $\Pi_{\mathcal{S}}$ is the orthogonal projection matrix on $\mathcal{S}$. Clearly Z_1 and Z_2 are Gaussian, and $\mathrm{Cov}(Z_1, Z_2) = \Pi_{\mathcal{S}} \Pi_{\mathcal{S}^\perp} = 0$, which implies Z_1 and Z_2 are independent. Choose $\mathcal{S} = \Delta^{-\frac{1}{2}} \mathrm{span}(\bar{\Gamma})$. Then $N_1 = \Delta^{\frac{1}{2}} \Pi_{\Delta^{-\frac{1}{2}} \mathrm{span}(\bar{\Gamma})} Z \in \mathrm{span}(\bar{\Gamma})$ while $N_2 = \Delta^{\frac{1}{2}} \Pi_{\Delta^{\frac{1}{2}} \mathrm{span}(\bar{\Gamma})^\perp} Z \in \Delta \mathrm{span}(\bar{\Gamma})^\perp$.

□

Proof of Proposition 1.2.8. By the factorization criterion for sufficiency ([33], p. 35, Theorem 6.5), the conditional density of $X | \left(\bar{\Gamma} S = s\right)$ has the form:

$$p(x|s) = \tilde{h}(\Gamma^T x, s) r(x),$$

for some functions $\tilde{h}$ and r. On the other hand, we know $X| \left(\bar{\Gamma}S = s\right) \sim \mathcal{N}(s, \Delta)$, thus we have the equality:

$$\tilde{h}(\Gamma^T x, s) r(x) = \phi_\Delta(x - s)$$

for all x and s. Plugging in $s = 0$ to both sides of the equality and rearranging we obtain $r(x) = \phi_\Delta(x)/\tilde{h}(\Gamma^T x, 0)$. To recover $p(x)$, the marginal density of X, we need to "integrate out" s according to the distribution of $\bar{\Gamma}S$, which yields:

$$p(x) = \int_{\mathbb{R}^p} \frac{\tilde{h}(\Gamma^T x, s)}{\tilde{h}(\Gamma^T x, s)} \phi_\Delta(x) dF(s),$$

where F is the distribution of $\bar{\Gamma}s$ (note that it cannot have a density with respect to Lebesgue measure on $\mathbb{R}^p$, since this vector is restricted to lie in a lower-dimensional linear subspace). Let $h(\Gamma^T x) = \int_{\mathbb{R}^p} \frac{\tilde{h}(\Gamma^T x, s)}{\tilde{h}(\Gamma^T x, s)} dF(s)$ to obtain the desired representation: $p(x) = h(\Gamma^T x)\phi_\Delta(x)$. Since $p(x)$ is a convolution of a non-Gaussian distribution with a Gaussian density, it must be differentiable. Note ϕ_Δ is differentiable as well. This shows that h is differentiable. □

Proof of Lemma 1.2.9. Let $\Gamma \in \mathbb{R}^{p \times d}$ have column space $\Delta^{-1}\text{span}(\bar{\Gamma})$. By elementary properties of Gaussian distributions, for all $s \in \text{span}(\bar{\Gamma})$:

$$\begin{bmatrix} X \\ \\ \Gamma^T X \end{bmatrix} \Bigg| \left(\bar{\Gamma}S = s\right) \sim \mathcal{N}\left(\begin{bmatrix} s \\ \\ \Gamma^T s \end{bmatrix}, \begin{bmatrix} \Delta & \Delta\Gamma \\ \\ \Gamma^T \Delta & \Gamma^T \Delta \Gamma \end{bmatrix} \right)$$

(if s is not a member of $\text{span}(\bar{\Gamma})$ then the probability of any event conditioned on $\{\text{span}(\bar{\Gamma}) = s\}$ will have probability 0). Therefore the distribution of $\left(X|\Gamma^T X = t, \bar{\Gamma}S = s\right)$ is also Gaussian. To check whether this distribution depends on s we just need to check whether the conditional mean $\mathbb{E}\left(X|\Gamma^T X = t, \bar{\Gamma}S = s\right)$ or the conditional covariance $\text{Cov}\left(X|\Gamma^T X = t, \bar{\Gamma}S = s\right)$ is constant (or not) for s. The conditional covariance does not depend on s; it is equal to:

$$\text{Cov}\left(X|\Gamma^T X = t, \bar{\Gamma}S = s\right) = \Delta - \Delta\Gamma\left(\Gamma^T \Delta\Gamma\right)^{-1}\Gamma^T\Delta.$$

So it just remains to check the conditional expectation:

$$\mathbb{E}\left(X|\Gamma^T X = t, \bar{\Gamma}S = s\right) = s + \mathbb{E}\left(N|\Gamma^T X = t, \bar{\Gamma}S = s\right).$$

Using $X = \bar{\Gamma}S + N$ we express the event $\{\Gamma^T X = t, \bar{\Gamma}S = s\}$ as the event $\{\Gamma^T N = t - \Gamma^T s, S = s\}$. Using a well-known formula for Gaussian regression coefficients we can compute $\mathbb{E}\left(N|\Gamma^T X = t, \bar{\Gamma}S = s\right)$, arriving at:

$$\mathbb{E}\left(X|\Gamma^T X = t, \bar{\Gamma}S = s\right) = s + \Delta\Gamma\left(\Gamma^T\Delta\Gamma\right)^{-1}\left(t - \Gamma^T s\right).$$

Clearly, if $\Delta\Gamma\left(\Gamma^T\Delta\Gamma\right)^{-1}\Gamma^T s = s$, then the conditional mean will not depend on s and the proof will be complete. We show that this is indeed the case. Some algebraic manipulations yield:

$$\begin{aligned}\Delta\Gamma\left(\Gamma^T\Delta\Gamma\right)^{-1}\Gamma^T s &= \Delta^{\frac{1}{2}}\left[\Delta^{\frac{1}{2}}\Gamma\left(\Gamma^T\Delta\Gamma\right)^{-1}\Gamma^T\Delta^{\frac{1}{2}}\right]\Delta^{-\frac{1}{2}}s\\ &= \Delta^{\frac{1}{2}}\left[\Pi_{\Delta^{\frac{1}{2}}\mathrm{span}(\Gamma)}\right]\Delta^{-\frac{1}{2}}s,\end{aligned}$$

where $\Pi_{\mathcal{S}}$ is the projection operator on the subspace $\mathcal{S}$. But $\mathrm{span}(\Gamma) = \Delta^{-1}\mathrm{span}(\bar{\Gamma})$, yielding:

$$\begin{aligned}\Delta^{\frac{1}{2}}\Pi_{\Delta^{-\frac{1}{2}}\mathrm{span}(\bar{\Gamma})}\Delta^{-\frac{1}{2}}s &= \Delta^{\frac{1}{2}}\Delta^{-\frac{1}{2}}s\\ &= s.\end{aligned}$$

□

Proof of Proposition 1.2.10. Using the representation of the density p given in Proposition 1.2.8 we see that

$$\nabla \log p(x) = \Gamma\frac{\nabla h(x)}{h(x)} + \Delta^{-1}x.$$

Replicate the arguments given in the proof of Proposition 1.2.5 using the above for $\nabla \log p(x)$ to complete the proof. □

Proof of Theorem 1.2.12. The equivalence of (ii) and (iii) is established in [37]. We will prove (i) and (ii) are equivalent.

(i)⇒(ii). First note that the $p \times p$ matrix $(\Gamma \quad \eta)$ is invertible. If it is not, then by Proposition 1.2.1 there exists $c \in \mathbb{R}^d$ such that $c^T V = k$ with probability 1. Hence, $c^T V \sim \mathcal{N}(k, 0)$. If C is any matrix with c as a row, then CV is a NGCA decomposition of V into independent Gaussian and non-Gaussian components. But no such decomposition exists by assumption. Therefore $(\Gamma \quad \eta)$ must be invertible. By Proposition 1.2.4 we can write

$$X = \bar{\Gamma}V' + \bar{\eta}G',$$

where $\bar{\Gamma} = \left(\Gamma\Gamma^T + \eta\eta^T\right)^{-1}\Gamma$, $\bar{\eta} = \left(\Gamma\Gamma^T + \eta\eta^T\right)^{-1}\eta$, $V' = V$ and $G' = G$ (see the proof of Proposition 1.2.4 for these equalities). Thus, the condition V does not have a d'-dimensional NGCA decomposition for $0 \leq d' < d$ is equivalent to the condition that there does not exist a full rank $d \times d$ matrix M such that the first coordinate of MV' has a marginal Gaussian distribution independent of the other $d-1$ coordinates.

(ii)⇒(i). By the proof of Proposition 1.2.4, we can choose Γ and η such that $\mathrm{span}(\Gamma) = \mathrm{span}(\bar{\eta})^\perp$ and $\mathrm{span}(\eta) = \mathrm{span}(\bar{\Gamma})^\perp$. Then,

$$\begin{bmatrix}\Gamma^T X \\ \eta^T X\end{bmatrix} = \begin{bmatrix}\Gamma^T \bar{\Gamma} V' \\ \eta^T \bar{\eta} G'\end{bmatrix}.$$

Therefore, $\Gamma^T X$ is a non-Gaussian random vector independent of the Gaussian vector $\eta^T X$, and we have obtained a NGCA decomposition as in Definition 1.1.1. Note that the $d \times d$ matrix $\Gamma^T \bar{\Gamma}$ must be invertible. If it is not, then we can find $v \in \mathbb{R}^d$ such that for all $d \times d$ matrices A, we have

$$0 = A\Gamma^T \bar{\Gamma} v = (\Gamma A)^T \bar{\Gamma} v.$$

Since $\operatorname{span}(\Gamma) = \operatorname{span}(\bar{\eta})^{\perp}$, we must have $\bar{\Gamma} v \in \operatorname{span}(\bar{\eta})$. This implies $\operatorname{span}(\bar{\Gamma}) \cap \operatorname{span}(\bar{\eta}) \neq \emptyset$, which implies the matrix $\begin{bmatrix}\bar{\Gamma} & \bar{\eta}\end{bmatrix}$ is not invertible: a contradiction. So $\Gamma^T \bar{\Gamma}$ is invertible.

Set $V = \Gamma^T \bar{\Gamma} V'$. Suppose there exists a d'-dimensional NGCA decomposition of V with $0 \leq d' < d$ and full-rank non-Gaussian and Gaussian spaces that do not intersect. Let $\Gamma_1 \in \mathbb{R}^{d \times d'}$ span the non-Gaussian space and $\eta_1 \in \mathbb{R}^{d \times (d-d')}$ span the Gaussian space, and set $V_1 = \Gamma_1^T V$ and $G_1 = \eta_1^T V$. By assumption, $[\Gamma_1 \quad \eta_1]$ is invertible. Without loss of generality we can assume the covariance matrix of the Gaussian component $G_1 = \eta_1^T V$ is diagonal (otherwise, we can transform G_1 by $G_1' = \operatorname{Cov}(G_1)^{-\frac{1}{2}} G_1 = \left(\eta_1^T \operatorname{Cov}(V) \eta_1\right)^{-\frac{1}{2}} \eta^T V$ and we would still have a NGCA decomposition with $\operatorname{Cov}(G_1') = I_{d-d'}$). By permuting the rows appropriately, this means the first component of $[\Gamma_1 \quad \eta_1]^T V = [\Gamma_1 \quad \eta_1]^T \Gamma^T \bar{\Gamma} V'$ has a marginal Gaussian distribution independent of the other coordinates. Since $[\Gamma_1 \quad \eta_1]^T \Gamma^T \bar{\Gamma}$ is invertible, this is a contradiction. □

Proof of Proposition 1.2.13. We have already shown $\Sigma \operatorname{span}(\eta) \subseteq \operatorname{span}(\Gamma)^{\perp}$. Since $\operatorname{span}(\eta)$ is full rank and Σ is invertible, the dimension of the subspace $\Sigma \operatorname{span}(\eta)$ must be $p - d$. The dimension of $\operatorname{span}(\Gamma)^{\perp}$ is also $p - d$. Therefore the two subspaces must be equal: $\Sigma \operatorname{span}(\eta) = \operatorname{span}(\Gamma)^{\perp}$.

We conclude $\operatorname{span}(\Gamma) = \Sigma^{-1}\operatorname{span}(\eta)$ by taking the orthogonal complements of both sides. □

Proof of Proposition 1.2.14. It is easy to check that $\mathbb{E}(\tilde{X}) = 0$ and $\operatorname{Cov}(\tilde{X}) = I_p$. Set $\tilde{\Gamma} = \Sigma^{1/2}\Gamma$ and $\tilde{\eta} = \Sigma^{1/2}\eta$. Then:

$$\begin{aligned}\begin{bmatrix}\tilde{\Gamma}^T \tilde{X} \\ \\ \tilde{\eta}^T \tilde{X}\end{bmatrix} &= \begin{bmatrix}\Gamma^T \Sigma^{\frac{1}{2}} \tilde{X} \\ \\ \eta^T \Sigma^{\frac{1}{2}} \tilde{X}\end{bmatrix} \\ &= \begin{bmatrix}V - \Gamma^T \mu \\ \\ G - \eta^T \mu\end{bmatrix},\end{aligned}$$

where V and G are the non-Gaussian and Gaussian components of X, respectively. Thus $\tilde{X}$ has a d-dimensional NGCA decomposition, with whitened non-Gaussian subspace span($\Sigma^{1/2}\Gamma$) and whitened Gaussian subspace span($\Sigma^{1/2}\eta$); these subspaces must be orthogonal by Proposition 1.2.13.

Since the non-Gaussian component of the decomposition of $\tilde{X}$ coincides with that of X up to an additive constant vector, if the decomposition of X is identifiable, it follows from Theorem 1.2.12 part (i) that the decomposition of $\tilde{X}$ is identifiable. □

Proof of Proposition 1.2.15. Let $\Sigma = \text{Cov}(X) \succ 0$ and $\mu = \mathbb{E}(X)$. Whiten X via the transformation $\tilde{X} = \Sigma^{-1/2}(X - \mu)$. Let $\tilde{\Gamma}$ span the whitened non-Gaussian space and $\tilde{\Gamma}_\perp$ span the whitened Gaussian space (these are the non-Gaussian and Gaussian spaces of $\tilde{X}$, which exist due to Proposition 1.2.14). Without loss of generality, assume $\tilde{\Gamma}$ and $\tilde{\Gamma}_\perp$ have orthogonal columns. We have for some non-Gaussian vector $\tilde{V}$ independent of Gaussian $\tilde{G}$:

$$\begin{bmatrix} \tilde{V} \\ \\ \tilde{G} \end{bmatrix} = \begin{bmatrix} \tilde{\Gamma}^T \tilde{X} \\ \\ \tilde{\Gamma}_\perp^T \tilde{X} \end{bmatrix}$$

Since $\tilde{X}$ is centered, we have $\mathbb{E}(\tilde{V}) = 0$ and $\mathbb{E}(\tilde{G}) = 0$. Computing the covariance matrix of both sides yields the equalities

$$\text{Cov}(\tilde{V}) = \tilde{\Gamma}^T \tilde{\Gamma} = I_d,$$

and

$$\text{Cov}(\tilde{G}) = \tilde{\Gamma}_\perp^T \tilde{\Gamma}_\perp = I_{p-d}.$$

Therefore pre-whitening makes $\tilde{G}$ a $p-d$-dimensional standard Gaussian, and makes $\tilde{V}$ have identity covariance structure. On the other hand, we have the relation:

$$\begin{bmatrix} \tilde{\Gamma}^T \tilde{X} \\ \\ \tilde{\Gamma}_\perp^T \tilde{X} \end{bmatrix} = \begin{bmatrix} \tilde{\Gamma}^T \Sigma^{-\frac{1}{2}}(X - \mu) \\ \\ \tilde{\eta}^T \Sigma^{-\frac{1}{2}}(X - \mu) \end{bmatrix}$$

If the non-Gaussian component $\tilde{V}$ has density f, then by the change of variable formula for multivariate distributions (see [6], p. 185) we obtain a new representation for the density $p(x)$ of X:

$$p(x) = (\det\Sigma)^{-\frac{1}{2}} f\left(\tilde{\Gamma}^T \Sigma^{-\frac{1}{2}}(x - \mu)\right) \phi_{p-d}\left(\tilde{\Gamma}_\perp^T \Sigma^{-\frac{1}{2}}(x - \mu)\right),$$

where ϕ_{p-d} is the density function of the standard normal distribution in $p - d$ dimensions. Recall that $\Sigma^{-1/2}\tilde{\Gamma}_\perp$ projects onto the Gaussian subspace of X. Since we assume the Gaussian component is centered, we must have $\Sigma^{-1/2}\tilde{\Gamma}_\perp \mu = 0$. Then we can write:

$$p(x) = (\det\Sigma)^{-\frac{1}{2}} \frac{f\left(\tilde{\Gamma}^T \Sigma^{-\frac{1}{2}}(x - \mu)\right)}{\phi_{p-d}\left(\tilde{\Gamma}^T \Sigma^{-\frac{1}{2}} x\right)} \phi_{p-d}\left(\tilde{\Gamma}_\perp^T \Sigma^{-\frac{1}{2}} x\right) \phi_{p-d}\left(\tilde{\Gamma}^T \Sigma^{-\frac{1}{2}} x\right).$$

Up to an additive constant $C(p,d)$ that only depends on p and d we have, from the form of the density of the standard normal distribution,

$$\begin{aligned}\log\phi_{p-d}\left(\tilde{\Gamma}_\perp^T\Sigma^{-\frac{1}{2}}x\right)+\log\phi_{p-d}\left(\tilde{\Gamma}^T\Sigma^{-\frac{1}{2}}x\right)&=-\frac{1}{2}\|\tilde{\Gamma}_\perp^T\Sigma^{-\frac{1}{2}}x\|_2^2-\frac{1}{2}\|\tilde{\Gamma}^T\Sigma^{-\frac{1}{2}}x\|_2^2+C(p,d)\\&=-\frac{1}{2}\|\Sigma^{-\frac{1}{2}}x\|_2^2+C(p,d),\end{aligned}$$

which follows from the fact that $\tilde{\Gamma}$ and $\tilde{\Gamma}_\perp$ form an orthonormal basis of $\mathbb{R}^p$. Thus, up to multiplicative constants that depend on p and d, $p(x)$ is equal to:

$$\frac{f\left(\tilde{\Gamma}^T\Sigma^{-\frac{1}{2}}(x-\mu)\right)}{\phi_{p-d}\left(\tilde{\Gamma}^T\Sigma^{-\frac{1}{2}}x\right)}\left(\det\Sigma\right)^{-\frac{1}{2}}\phi_p\left(\Sigma^{-\frac{1}{2}}x\right).$$

Clearly $\left(\det\Sigma\right)^{-\frac{1}{2}}\phi_p\left(\Sigma^{-\frac{1}{2}p}x\right)=\phi_\Sigma(x)$. Recall $\Sigma^{-1/2}\tilde{\Gamma}$ spans the non-Gaussian subspace of X. We can replace it with Γ. Then set:

$$q(\Gamma^Tx)=\frac{f\left(\Gamma^T(x-\mu)\right)}{\phi_{p-d}\left(\Gamma^Tx\right)},$$

absorbing any leftover multiplicative constants. Note that q is differentiable in x since it is the quotient of two differentiable functions (f is assumed differentiable in the theorem) and $\phi_{p-d}>0$. Thus $p(x)$ has the desired form. □

Proof of Proposition 1.2.16. Using the representation of the density p given in Proposition 1.2.15 we see that

$$\nabla\log p(x)=\Gamma\frac{\nabla q(x)}{q(x)}+\Sigma^{-1}x.$$

Replicate the arguments given in the proof of Proposition 1.2.5 using the above for $\nabla\log p(x)$ to complete the proof. □

A.2 Proofs from Section 1.4

Proof of Proposition 1.4.1. Let $\mathcal{O}$ be a $p\times p$ orthogonal matrix represented in block form by $\mathcal{O}=[\Gamma\quad\Gamma_\perp]^T$ (here Γ is $p\times d$). Blockwise we have, for each k,

$$\mathcal{O}M_k\mathcal{O}^T=\begin{bmatrix}\Gamma^TM_k\Gamma & \Gamma^TM_k\Gamma_\perp\\ \Gamma_\perp^TM_k\Gamma & \Gamma_\perp^TM_k\Gamma_\perp\end{bmatrix}.$$

Hence,

$$\|\mathcal{O}M_k\mathcal{O}^T\|_F^2 = \|\Gamma^T M_k\Gamma\|_F^2 + \|\Gamma^T M_k\Gamma_\perp\|_F^2 + \|\Gamma_\perp^T M_k\Gamma\|_F^2 + \|\Gamma_\perp^T M_k\Gamma_\perp\|_F^2.$$

On the other hand, the Frobenius norm is invariant under orthogonal transformations: $\|\mathcal{O}M_k\mathcal{O}^T\|_F^2 = \|M_k\|_F^2$. Therefore, we have the inequality

$$\|\Gamma^T M_k\Gamma\|_F^2 \leq \|M_k\|_F^2$$

with equality if and only if $\Gamma^T M_k\Gamma_\perp = 0$, $\Gamma_\perp^T M_k\Gamma = 0$ and $\Gamma_\perp^T M_k\Gamma_\perp$. But this occurs precisely when $\mathcal{O} = \mathcal{O}_0$ and $\Gamma = \Gamma_0$. That is,

$$\|\Gamma_0^T M_k\Gamma_0\|_F^2 = \|M_k\|_F^2.$$

Since we assume the diagonalization holds for each k it follows that

$$\sum_{k=1}^{K}\|\Gamma^T M_k\Gamma\|_F^2 \leq \sum_{k=1}^{K}\|\Gamma_0^T M_k\Gamma_0\|_F^2.$$

Finally, using the invariance of the Frobenius norm to orthogonal transformations, we conclude

$$Q(\Gamma) \leq Q(\Gamma_0 U)$$

for all $p \times d$ orthogonal matrices Γ and all $d \times d$ orthogonal matrices $\mathcal{U}$. □

Appendix B

Appendix for Chapter 2

B.1 Uniform bounds for consistency proofs.

This section contains proofs of the uniform bounds of Lemmas 2.3.1 and 2.3.6.

Proof of Lemma 2.3.1. From Proposition B.3.1, to get our desired result it suffices to show:

$$\sup_{\Gamma\in\mathfrak{G}_{d,p}} \iint \phi_d(s)\phi_{p-d}(t)\big|\mathcal{X}(\Gamma s+\Gamma_\perp t;\widehat{P}_n)-\mathcal{X}(\Gamma s+\Gamma_\perp t;P)\big|\mathrm{d}s\mathrm{d}t = o_{P^*}(1),$$

and:

$$\sup_{\Gamma\in\mathfrak{G}_{d,p}} \int \phi_d(s)\big|\mathcal{X}(\Gamma s;P)-\mathcal{X}(\Gamma s;\widehat{P}_n)\big|\mathrm{d}s = o_{P^*}(1).$$

We work in outer probability to avoid measurability issues that may arise from taking the supremum of an uncountable collection of random variables (side note: another approach is to assume the stochastic process $\rho(\Gamma,\widehat{P}_n)$ is separable). Fix any $r>0$. Then:

$$\begin{aligned}
&\iint \phi_d(s)\phi_{p-d}(t)\big|\mathcal{X}(\Gamma s+\Gamma_\perp t;\widehat{P}_n)-\mathcal{X}(\Gamma s+\Gamma_\perp t;P)\big|\mathrm{d}s\mathrm{d}t\\
&\leq \iint_{\|(s,t)\|_2\leq r} \phi_d(s)\phi_{p-d}(t)\big|\mathcal{X}(\Gamma s+\Gamma_\perp t;\widehat{P}_n)-\mathcal{X}(\Gamma s+\Gamma_\perp t;P)\big|\mathrm{d}s\mathrm{d}t\\
&+2\int_{\|(s,t)\|_2>r} \phi_d(s)\phi_{p-d}(t)\mathrm{d}s\mathrm{d}t,
\end{aligned}$$

where we use the bound $\big|\mathcal{X}(\Gamma s+\Gamma_\perp t;\widehat{P}_n)-\mathcal{X}(\Gamma s+\Gamma_\perp t;P)\big|\leq 2$. We bound both terms. Clearly,

$$2\int_{\|(s,t)\|_2>r}\phi_d(s)\phi_{p-d}(t)\mathrm{d}s\mathrm{d}t = 2\int_{\|u\|_2>r}\phi_p(u)\mathrm{d}u$$
$$= 2\mathbb{P}\left(\chi_p^2 \geq r^2\right),$$

where χ_p^2 is a generic Chi-square random variable on p degrees of freedom. For the other term,

$$\iint_{\|(s,t)\|_2\leq r}\phi_d(s)\phi_{p-d}(t)\big|\mathcal{X}(\Gamma s+\Gamma_\perp t;\widehat{P}_n)-\mathcal{X}(\Gamma s+\Gamma_\perp t;P)\big|\mathrm{d}s\mathrm{d}t$$
$$\leq \sup_{\|(s,t)\|_2\leq r}\big|\mathcal{X}(\Gamma s+\Gamma_\perp t;\widehat{P}_n)-\mathcal{X}(\Gamma s+\Gamma_\perp t;P)\big|.$$

Note that the transformation $(s,t)\to\Gamma s+\Gamma_\perp t$ is full-rank and isometric with respect to the Euclidean metric; that is, $\|(s,t)\|_2=\|\Gamma s+\Gamma_\perp t\|_2$. This implies:

$$\sup_{\|(s,t)\|_2\leq r}\big|\mathcal{X}(\Gamma s+\Gamma_\perp t;\widehat{P}_n)-\mathcal{X}(\Gamma s+\Gamma_\perp t;P)\big| = \sup_{\|u\|_2\leq r}\big|\mathcal{X}(u;\widehat{P}_n)-\mathcal{X}(u,P)\big|.$$

Note how the term on the right hand side does not depend on Γ or $\Gamma_\perp$. We have therefore obtained the bound:

$$\sup_{\Gamma\in\mathfrak{G}_{d,p}}\iint\phi_d(s)\phi_{p-d}(t)\big|\mathcal{X}(\Gamma s+\Gamma_\perp t;\widehat{P}_n)-\mathcal{X}(\Gamma s+\Gamma_\perp t;P)\big|\mathrm{d}s\mathrm{d}t$$
$$\leq \sup_{u\in\mathbb{R}^p:\|u\|_2\leq r}\big|\mathcal{X}(u;\widehat{P}_n)-\mathcal{X}(u;P)\big| + 2\mathbb{P}\left(\chi_p^2\geq r^2\right).$$

By Theorem 2.1 in [14], given any fixed r, $\sup_{u\in\mathbb{R}^p:\|u\|_2\leq r}\big|\mathcal{X}(u;\widehat{P}_n)-\mathcal{X}(u;P)\big| = o_P(1)$ (in [14] measurability difficulties are obviated by assuming the stochastic process is separable). Hence, for any $\epsilon>0$, choose r large enough so that $2\mathbb{P}\left(\chi_p^2\geq r^2\right)\leq\epsilon$; this suffices to show the right hand side of the above display is $o_P(1)$.

To show

$$\sup_{\Gamma\in\mathfrak{G}_{d,p}}\int\phi_d(s)\big|\mathcal{X}(\Gamma s;P)-\mathcal{X}(\Gamma s;\widehat{P}_n)\big|\mathrm{d}s = o_P(1),$$

observe the fact that for any $s\in\mathbb{R}^d$ and orthonormal $\Gamma\in\mathbb{R}^{p\times d}$,
$\|\Gamma s\|_2^2 = s^T\Gamma^T\Gamma s = s^Ts = \|s\|_2^2$; we can apply the preceding arguments for s instead of (s,t).
□

We now prove Lemma 2.3.6 from page 43.

Proof of Lemma 2.3.6. From Proposition B.3.1:

$$\begin{aligned}
&\big|\rho(\Gamma, P_1) - \rho(\Gamma, P_2)\big| \\
&\qquad \leq 4 \int \phi_d(s)\phi_{p-d}(t) \Big\{ \big|\mathcal{X}(\Gamma s + \Gamma_\perp t; \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})) - \mathcal{X}(\Gamma s + \Gamma_\perp t; \widehat{\tilde{P}}_n(\Sigma, \mu))\big| \\
&\qquad\qquad + \big|\mathcal{X}(\Gamma s; \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})) - \mathcal{X}(\Gamma s; \widehat{\tilde{P}}_n(\Sigma, \mu))\big| \Big\} \mathrm{d}s\mathrm{d}t.
\end{aligned}$$

Therefore, if

$$\sup_{\Gamma \in \mathfrak{G}_{d,p}} \iint \phi_d(s)\phi_{p-d}(t) \big|\mathcal{X}(\Gamma s + \Gamma_\perp t; \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})) - \mathcal{X}(\Gamma s + \Gamma_\perp t; \widehat{P}_n(\Sigma, \mu))\big| \mathrm{d}s\mathrm{d}t = o_{P^*}(1),$$

and if

$$\sup_{\Gamma \in \mathfrak{G}_{d,p}} \int \phi_d(s) \big|\mathcal{X}(\Gamma s; \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})) - \mathcal{X}(\Gamma s; \widehat{\tilde{P}}_n(\Sigma, \mu))\big| \mathrm{d}s = o_{P^*}(1).$$

then the lemma holds. Both are proved in the exact same way: we will prove the first.

We have:

$$\begin{aligned}
&\big|\mathcal{X}(\Gamma s + \Gamma_\perp t; \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})) - \mathcal{X}(\Gamma s + \Gamma_\perp t; \widehat{\tilde{P}}_n(\Sigma, \mu))\big| \\
\leq & \frac{1}{n} \sum_{i=1}^n \left| \exp\left(i(\Gamma s + \Gamma_\perp t)^T \hat{X}_i \right) - \exp\left(i(\Gamma s + \Gamma_\perp t)^T \tilde{X}_i \right) \right|,
\end{aligned}$$

For any real numbers a and b, $\big|\exp(ia) - \exp(ib)\big| = \big|\int_a^b \exp(ix)\mathrm{d}x\big| \leq |a-b|$ since $\big|\exp(ix)\big| \leq 1$ for all x. Hence:

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \big| \exp(i(\Gamma s + \Gamma_\perp t)^T \hat{X}_i) - \exp(i(\Gamma s + \Gamma_\perp t)^T \tilde{X}_i) \big| \\
\leq & \frac{1}{n} \sum_{i=1}^n \big| \left(\Gamma s + \Gamma_\perp t\right)^T \left(\hat{X}_i - \tilde{X}_i \right) \big| \\
\leq & \frac{1}{n} \sum_{i=1}^n \|\Gamma s + \Gamma_\perp t\|_2 \|\hat{X}_i - \tilde{X}_i\|_2 \\
= & \frac{1}{n} \sum_{i=1}^n \|(s, t)\|_2 \|\hat{X}_i - \tilde{X}_i\|_2,
\end{aligned}$$

where we applied the equality $\|\Gamma s + \Gamma_\perp t\|_2 = \|(s,t)^T\|_2$; note that the parameters Γ and $\Gamma_\perp$ have dropped out of the expression. At this step plug in $\hat{X}_i = \hat{\Sigma}^{-1/2}(X_i - \hat{\mu})$ and $\tilde{X}_i = \Sigma^{-1/2}(X_i - \mu)$; the above is equal to:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\|(s,t)^T\|_2\|(\widehat{\Sigma}^{-\frac{1}{2}} - \Sigma^{-\frac{1}{2}})\,(X_i - \mu) - \widehat{\Sigma}^{-\frac{1}{2}}\,(\widehat{\mu} - \mu)\|_2 \\
\leq&\frac{1}{n}\sum_{i=1}^{n}\|(s,t)^T\|_2\left\{\left\|\widehat{\Sigma}^{-\frac{1}{2}} - \Sigma^{-\frac{1}{2}}\right\|_2\|X_i - \mu\|_2 + \left\|\widehat{\Sigma}^{-\frac{1}{2}}\right\|_2\|\widehat{\mu} - \mu\|_2\right\},
\end{aligned}$$

where the last line follows from the Cauchy-Schwarz inequality (here $\|\cdot\|_2$, when applied to a matrix, refers to the usual operator (spectral) norm). Next we take the integral over s and t:

$$\begin{aligned}
&\sup_{\Gamma \in \mathfrak{G}_{d,p}} \int \phi_d(s)\phi_{p-d}(t)\left|\mathcal{X}(\Gamma s + \Gamma_\perp t; \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})) - \mathcal{X}(\Gamma s + \Gamma_\perp t; \widehat{P}_n(\Sigma, \mu))\right|\mathrm{d}s\mathrm{d}t \\
&\qquad \leq \mathbb{E}\left(\chi_p\right)\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\|X_i - \mu\|_2\right)\|\widehat{\Sigma}^{-\frac{1}{2}} - \Sigma^{-\frac{1}{2}}\|_2 + \|\widehat{\Sigma}^{-\frac{1}{2}}\|_2\|\widehat{\mu} - \mu\|_2\right\},
\end{aligned}$$

where χ_p has a Chi distribution on p degrees of freedom. Note that $\mathbb{E}(\chi_p) \leq \sqrt{p}$. We show the other terms are $o_P(1)$. By the weak law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n}\|X_i - \mu\|_2 \xrightarrow{P} \mathbb{E}\|X_1 - \mu\|_2 < \infty,$$

where the last equality holds since P by assumption has finite second moments. Consistency of $\hat{\Sigma}$ and Σ positive definite imply $\hat{\Sigma}^{-1}$ exists with probability tending to 1; continuity of the mapping $\Sigma \to \Sigma^{-1/2}$ with respect to the operator norm implies:

$$\|\widehat{\Sigma}^{-\frac{1}{2}} - \Sigma^{-\frac{1}{2}}\|_2 \xrightarrow{P} 0,$$

and:

$$\|\hat{\Sigma}^{-1/2}\|_2 \xrightarrow{P} \|\Sigma^{-1/2}\|_2.$$

By the law of large numbers:

$$\|\widehat{\mu} - \mu\|_2 \xrightarrow{P} 0.$$

Therefore:

$$\mathbb{E}\left(\chi_p\right)\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\|X_i - \mu\|_2\right)\|\widehat{\Sigma}^{-\frac{1}{2}} - \Sigma^{-\frac{1}{2}}\|_2 + \|\widehat{\Sigma}^{-\frac{1}{2}}\|_2\|\widehat{\mu} - \mu\|_2\right\} \xrightarrow{P} 0,$$

which proves the lemma.

□

B.2 Asymptotic normality proofs.

Proof of Theorem 2.3.4. Recall the mapping $\Gamma(B)$ defined on $\mathbb{R}^{(p-d)\times d}$ given by:

$$\Gamma(B) = (\Gamma_0 \;\; \Gamma_{0_\perp}) \exp\left(\begin{bmatrix} 0 & -B^\top \\ B & 0 \end{bmatrix} \right) J_{p,d},$$

where $J \in \mathbb{R}^{p\times d}$ consists of the first d columns of the $p \times p$ identity matrix. Since the matrix exponential map is smooth, $\Gamma(B)$ is smooth. Moreover, the function $r(x, y, \Gamma)$ in Proposition 2.3.3 on page 41 is obviously smooth, which implies the function of B formed by the composition $\rho(\Gamma(B), \widehat{P}_n) = \iint r(x, y, \Gamma(B))\mathrm{d}\widehat{P}_n(x)\mathrm{d}\widehat{P}_n(y)$ is smooth. $\widehat{B}_n$ therefore satisfies a stationary condition: since $\widehat{\Gamma}_n = \Gamma(\widehat{B}_n)$, $\widehat{B}_n$ is a global minimizer of $\rho(\Gamma(B), \widehat{P}_n)$, and therefore:

$$0 = \nabla_B \rho(\Gamma(\widehat{B}_n), \widehat{P}_n).$$

The idea of our proof is to use the smoothness introduced by the mapping $\Gamma(B)$ to take the Taylor expansion of the above quantity about 0 (recall $\Gamma(0) = \Gamma_0$, the true non-Gaussian subspace parameter). To ease the notation we vectorize derivatives by stacking columns; this is accomplished by identifying B with $\mathrm{vec}(B)$:

$$\nabla_{\mathrm{vec}(B)} \rho(\Gamma(B), \widehat{P}_n) = \mathrm{vec}\left(\nabla_B \rho(\Gamma(B), \widehat{P}_n) \right).$$

The notation $\nabla_{\mathrm{vec}(B)}$ indicates that the derivatives should be placed in a $d(p-d)$-dimensional vector by stacking the columns of the matrix variable B according to $\mathrm{vec}(B)$. Taking the Taylor expansion yields:

$$\begin{aligned} 0 &= \nabla_{\mathrm{vec}(B)} \rho(\Gamma(\widehat{B}_n), \widehat{P}_n) && \text{(B.1)} \\ &= \nabla_{\mathrm{vec}(B)} \rho(\Gamma_0, \widehat{P}_n) + \nabla^2_{\mathrm{vec}(B)} \rho(\Gamma_0, \widehat{P}_n)\mathrm{vec}(\widehat{B}_n) + R(\widehat{B}_n, \widehat{P}_n), && \text{(B.2)} \end{aligned}$$

where $\nabla^2_{\mathrm{vec}(B)} \rho(\Gamma(B), P)$ is a $d(p-d) \times d(p-d)$ Hessian matrix of $\rho(\Gamma(B), P)$ with respect to B. The term $R(\widehat{B}_n, \widehat{P}_n)$ is the remainder. To analyze (B.1) we will obtain the asymptotic behavior of the terms $R(\widehat{B}_n, \widehat{P}_n)$, $\nabla^2_{\mathrm{vec}(B)} \rho(\Gamma_0, \widehat{P}_n)\mathrm{vec}(\widehat{B}_n)$, and $\nabla_{\mathrm{vec}(B)} \rho(\Gamma_0, \widehat{P}_n)$ in that order.

Show remainder $R(\widehat{B}_n, \widehat{P}_n) = O_P\left(\|\widehat{B}_n\|_F^2 \right)$.

To begin, observe that $R(\widehat{B}_n, \widehat{P}_n)$ is a $d(p-d)$ dimensional vector whose jth entry can be written using Lagrange's remainder theorem for multivariable functions as

$$R(\widehat{B}_n, \widehat{P}_n) = \mathrm{vec}(\widehat{B}_n)^T \left[\int_0^1 (1-\xi)\nabla^2_{\mathrm{vec}(B)} \frac{\partial}{\partial B_j}\rho(\Gamma(\xi\widehat{B}_n), \widehat{P}_n)\mathrm{d}\xi\right] \mathrm{vec}(\widehat{B}_n),$$

where $\nabla^2_{\mathrm{vec}(B)} \frac{\partial}{\partial B_j}\rho(\Gamma(\xi\widehat{B}_n), \widehat{P}_n)$ is a $d(p-d) \times d(p-d)$ matrix whose kl element is equal to:

$$\frac{\partial^3}{\partial B_k \partial B_l \partial B_j}\rho(\Gamma(\xi\widehat{B}_n), \widehat{P}_n);$$

By Lemma B.3.2 we know that for any B we have $\left|\frac{\partial^3}{\partial B_j \partial B_k \partial B_l} r(x, y, \Gamma(B))\right| \le K(p, d)\,[\|x\|_2^3 + \|y\|_2^3]$, where K is a constant that only depends on p and d, which are held fixed as $n \to \infty$. Since $\rho(\Gamma(B), \widehat{P}_n) = \iint \mathrm{d}P(x)\mathrm{d}P(y) r(x, y, \Gamma(B))$ we have $\left|\frac{\partial^3}{\partial B_j \partial B_k \partial B_l}\rho(\Gamma(B), \widehat{P}_n)\right| \le K_1(p, d)\frac{1}{n}\sum_{i=1}^n \|X_i\|_2^3$. This yields the bound:

$$\begin{aligned} &\left|\mathrm{vec}(\widehat{B}_n)^T \left[\int_0^1 (1-\xi)\nabla^2_{\mathrm{vec}(B)} \frac{\partial}{\partial B_j}\rho(\xi\widehat{B}_n, \widehat{P}_n)\mathrm{d}\xi\right] \mathrm{vec}(\widehat{B}_n)\right| \\ &\le \left[\int_0^1 (1-\xi)\left\|\nabla^2_{\mathrm{vec}(B)} \frac{\partial}{\partial B_j}\rho(\xi\widehat{B}_n, \widehat{P}_n)\right\|_F \mathrm{d}\xi\right] \|\mathrm{vec}(\widehat{B}_n)\|_2^2 \\ &\le K_1(p, d)\frac{d^2(p-d)^2}{2}\left[\frac{1}{n}\sum_{i=1}^n \|X_i\|^3\right] \|\widehat{B}_n\|_F^2. \end{aligned}$$

It follows that $\|R(\widehat{B}_n, \widehat{P}_n)\|_2 \le K'(p, d)\,[n^{-1}\sum_i \|X\|_i^3]\,\|\widehat{B}_n\|_F^2$ for some constant $K'(p, d)$. P_0 is assumed to have finite third moments, therefore $n^{-1}\sum_i \|X_i\|^3 = O_P(1)$ as $n \to \infty$. We conclude:

$$\boxed{R(\widehat{B}_n, \widehat{P}_n) = O_P(\|\widehat{B}_n\|_F^2).}$$

Show $\nabla^2_{\mathrm{vec}(B)}\rho(\Gamma_0, \widehat{P}_n)\mathrm{vec}(\widehat{B}_n) \approx \frac{2}{3}\left(\frac{1}{3}\right)^{\frac{p-d}{2}}[M(F) \otimes I_{p-d}]\,\mathrm{vec}(\widehat{B}_n)$**.**

From Proposition 2.3.3 we can write $\nabla^2_{\mathrm{vec}(B)}\rho(\Gamma_0, \widehat{P}_n)$ as a V-statistic with a matrix-valued kernel:

$$\frac{1}{n^2}\sum_{i,j}\nabla^2_{\mathrm{vec}(B)} r(X_i, X_j, \Gamma_0).$$

By Lemma B.3.2, $\left|\frac{\partial^2}{\partial B_k \partial B_l} r(x, y, \Gamma(B))\right| \le K(p, d)(\|x\|_2^2 + \|y\|_2^2)$ where $K(p, d)$ is a constant that only depends on p and d. First, we remove the diagonal term: since the third moments of P_0 are assumed finite, we have $\mathbb{E}_{P_0}\left(\frac{\partial^2}{\partial B_k \partial B_l} r(X, X, \Gamma(B))\right) < \infty$, which in turn implies, by the weak law of large numbers,

$$\frac{1}{n^2}\sum_{i=1}^{n}\frac{\partial^2}{\partial B_k \partial B_l} r(X_i, X_i, \Gamma_0) = O_P(n^{-1}).$$

This is true for all indices $k, l = 1, \ldots, d(p-d)$. Since the dimension of the matrix is bounded with n we can use a simple union bound to obtain:

$$\left\| \frac{1}{n^2}\sum_{i=1}^{n} \nabla^2_{\mathrm{vec}(B)} r(X_i, X_i, \Gamma_0) \right\|_F = O_P(n^{-1}).$$

This matrix multiplies $\mathrm{vec}(\widehat{B}_n)$, thereby contributing a term of order $O_P(n^{-1}\|\widehat{B}_n\|_F)$ to the asymptotic expansion.
For the off-diagonal terms, write:

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i\neq j} \nabla^2_{\mathrm{vec}(B)} r(X_i, X_j, \Gamma_0)\mathrm{vec}(\widehat{B}_n) = \\
&\mathbb{E}_{P_0}\left(\nabla^2_{\mathrm{vec}(B)} r(X, Y, \Gamma_0)\right)\mathrm{vec}(\widehat{B}_n) \\
&+ \left[\frac{1}{n^2}\sum_{i\neq j} \nabla^2_{\mathrm{vec}(B)} r(X_i, X_j, \Gamma_0) - \mathbb{E}_{P_0}\left(\nabla^2_{\mathrm{vec}(B)} r(X, Y, \Gamma_0)\right)\right]\mathrm{vec}(\widehat{B}_n).
\end{aligned}$$

Each entry of the random matrix

$$\frac{1}{n^2}\sum_{i\neq j} \nabla^2_{\mathrm{vec}(B)} r(X_i, X_j, \Gamma_0) - \mathbb{E}_{P_0}\left(\nabla^2_{\mathrm{vec}(B)} r(X, Y, \Gamma_0)\right)$$

is a U-statistic with an integrable, mean zero kernel. By the law of large numbers for U-statistics [24], each entry is of order $o_P(1)$. Being of fixed dimension, we can use the union bound to assert the matrix as a whole is $o_P(1)$. Hence

$$\left[\frac{1}{n^2}\sum_{i\neq j} \nabla^2_{\mathrm{vec}(B)} r(X_i, X_j, \Gamma_0) - \mathbb{E}_{P_0}\left(\nabla^2_{\mathrm{vec}(B)} r(X, Y, \Gamma_0)\right)\right]\mathrm{vec}(\widehat{B}_n)$$

is of order $o_P(\|\widehat{B}_n\|_F)$.

Having dealt with the remainder, we now calculate $\mathbb{E}_{P_0}\left(\nabla^2_{\mathrm{vec}(B)} r(X, Y, \Gamma_0)\right)\mathrm{vec}(\widehat{B}_n)$ directly. We can write

$$\begin{aligned}
\mathbb{E}_{P_0}\left(\nabla^2_{\mathrm{vec}(B)} r(X, Y, \Gamma_0)\right) &= \iint \mathrm{d}P_0(x)\mathrm{d}P_0(y)\nabla^2_{\mathrm{vec}(B)} r(x, y, \Gamma_0) \\
&= \nabla^2_{\mathrm{vec}(B)} \rho(\Gamma_0, P_0).
\end{aligned}$$

Let $D(s,t,B,P) = \mathcal{X}\left(\Gamma(B)s + \Gamma_\perp(B)t; P\right) - e^{\|t\|_2^2/2}\mathcal{X}\left(\Gamma(B)s; P\right)$. Then:

$$\rho(\Gamma(B),P) = \iint D(s,t,B,P)\overline{D(s,t,B,)}\phi_d(s)\phi_{p-d}(t)\mathrm{d}s\mathrm{d}t,$$

where $\overline{z}$ denotes the complex conjugate of $z \in \mathbb{C}$. The Hessian of ρ in B is given, in terms of $D(s,t,B,P)$, by:

$$\nabla^2_{\mathrm{vec}(B)}\rho(\Gamma(B),P) = 2\iint \Big[\nabla^2_{\mathrm{vec}(B)}D(s,t,B,P)\,\overline{D(s,t,B,P)} + \nabla_{\mathrm{vec}(B)}D(s,t,B,P)\nabla_{\mathrm{vec}(B)}\overline{D(s,t,B,P)}^T\Big]\phi_d(s)\phi_{p-d}(t)\mathrm{d}s\mathrm{d}t.$$

Clearly $D(s,t,0,P_0) = 0$ for all s and t since $\Gamma(0) = \Gamma_0$ is the non-Gaussian subspace parameter of P_0. Therefore:

$$\nabla^2_{\mathrm{vec}(B)}\rho(\Gamma_0,P_0) = 2\iint \nabla_{\mathrm{vec}(B)}D(s,t,0,P_0)\nabla_{\mathrm{vec}(B)}\overline{D(s,t,0,P_0)}^T\phi_d(s)\phi_{p-d}(t)\mathrm{d}s\mathrm{d}t.$$

To obtain an explicit expression for the Hessian we need to compute $\nabla_{\mathrm{vec}(B)}D(s,t,B,P)$ and evaluate at $B = 0$ and $P = P_0$. Write

$$D(s,t,B,P) = \mathbb{E}_P\left[\exp\{iX^T(\Gamma(B)s + \Gamma_\perp(B)t)\}\right] - e^{-\|t\|_2^2/2}\mathbb{E}_P\left[\exp\{iX^T(\Gamma(B)s)\}\right].$$

If P has a finite first moment, then its characteristic function is continuously differentiable, and we can exchange the derivative and the expectation operator. This is true for $P = P_0$ by assumption. This allows us to differentiate inside the expectation operator:

$$\begin{aligned}
&\nabla_{\mathrm{vec}(B)}\exp\left(ix^T[\Gamma(B)s + \Gamma_\perp(B)t]\right)\\
&= \mathrm{vec}\left[\nabla_B \exp\left(ix^T[\Gamma(B)s + \Gamma_\perp(B)t]\right)\right]\\
&= \mathrm{vec}\left[\exp\left(ix^T(\Gamma(B)s + \Gamma_\perp(B)t)\right)i\nabla_B\left(x^T[\Gamma(B)s + \Gamma_\perp(B)t]\right)\right].
\end{aligned}$$

We need to calculate $\nabla_B\ [x^T\Gamma(B)s]$ and $\nabla_B\ [x^T\Gamma_\perp(B)t]$ and evaluate at $B = 0$. To do so, consider first the more general problem of computing $\nabla_B f(\Gamma(B))$ and $\nabla_B g(\Gamma_\perp(B))$ for real-valued differentiable functions f and g. Our strategy will be element-wise: compute $\frac{\partial}{\partial B_{jk}}f(\Gamma(B))$ and $\frac{\partial}{\partial B_{jk}}g(\Gamma_\perp(B))$ for $j = 1,\ldots,(p-d)$, $k = 1,\ldots,d$. By the chain rule,

$$\frac{\partial}{\partial B_{jk}} f(\Gamma(B)) = \operatorname{Tr}\left[\left(\frac{\partial}{\partial B_{jk}}\Gamma(B)\right)^T \nabla f(\Gamma(B))\right],$$

and

$$\frac{\partial}{\partial B_{jk}} g(\Gamma_\perp(B)) = \operatorname{Tr}\left[\left(\nabla g(\Gamma_\perp(B))\right)^T \frac{\partial}{\partial B_{jk}}\Gamma_\perp(B)\right],$$

where Tr denotes the usual trace operator. We simultaneously obtain expressions for $\frac{\partial}{\partial B_{jk}}\Gamma(B)$ and $\frac{\partial}{\partial B_{jk}}\Gamma_\perp(B)$ by considering the full $p \times p$ orthogonal matrix $(\Gamma(B) \;\; \Gamma_\perp(B))$ and taking the derivative:

$$\frac{\partial}{\partial B_{jk}}\left(\Gamma(B) \;\; \Gamma_\perp(B)\right) = \left(\Gamma_{0_\perp}\frac{\partial}{\partial B_{jk}}B \quad -\Gamma_0\frac{\partial}{\partial B_{jk}}B^T\right)\exp\left(\begin{bmatrix} 0 & -B^T \\ B & 0 \end{bmatrix}\right).$$

$\frac{\partial}{\partial B_{jk}}\Gamma(B)$ is given by the first d columns of the matrix in the above display; $\frac{\partial}{\partial B_{jk}}\Gamma_\perp(B)$ consists of the remaining $p-d$ columns. Note that $\frac{\partial}{\partial B_{jk}}B$ is a $(p-d)\times d$ matrix with the jk entry equal to 1 and all other entries equal to 0. We denote it by $\mathbf{1}_{jk}^{(p-d)\times d}$. Hence:

$$\begin{aligned}\left.\frac{\partial}{\partial B_{jk}} f(\Gamma(B))\right|_{B=0} &= \operatorname{Tr}\left[\left(\mathbf{1}_{jk}^{(p-d)\times d}\right)^T \Gamma_{0_\perp}^T \nabla f(\Gamma_0)\right] \\ &= \left(\Gamma_{0_\perp}^T \nabla f(\Gamma_0)\right)_{jk},\end{aligned}$$

since the matrix $\left(\mathbf{1}_{jk}^{(p-d)\times d}\right)^T$ picks out the jk element of the $(p-d)\times d$ matrix $\Gamma_{0_\perp}^T \nabla f(\Gamma_0)$. Similarly, using the cycle property of the trace operator,

$$\begin{aligned}\left.\frac{\partial}{\partial B_{jk}} g(\Gamma_\perp(B))\right|_{B=0} &= \operatorname{Tr}\left[\left(\nabla g(\Gamma_{0_\perp})\right)^T \left(-\Gamma_0\left(\mathbf{1}_{jk}^{(p-d)\times d}\right)^T\right)\right] \\ &= -\operatorname{Tr}\left[\left(\mathbf{1}_{jk}^{(p-d)\times d}\right)^T \left(\nabla g(\Gamma_{0_\perp})^T \Gamma_0\right)\right] \\ &= -\left(\nabla g(\Gamma_{0_\perp})^T\Gamma_0\right)_{jk}.\end{aligned}$$

We have derived general formulas:

$$\nabla_B f(\Gamma(B))\big|_{B=0} = \Gamma_{0_\perp}^T \nabla f(\Gamma_0)$$

and

$$\nabla_B \; g(\Gamma_\perp(B))\big|_{B=0} = -\nabla g(\Gamma_{0_\perp})^T\Gamma_0.$$

To carry on our derivation, we apply them to the specific functions at hand:

$$\nabla_B \left[x^T\Gamma(B)s]\right|_{B=0} = \Gamma_{0_\perp}^T x s^T$$

and

$$\nabla_B \left[x^T\Gamma_\perp(B)t]\right|_{B=0} = -tx^T\Gamma_0.$$

The result of these calculations yields:

$$\nabla_{\text{vec}(B)} \exp\left(ix^T[\Gamma(B)s + \Gamma_\perp(B)t]\right)\Big|_{B=0} \\ = \exp\left(ix^T[\Gamma_0 s + \Gamma_{0_\perp} t]\right) i\text{vec}\left(\Gamma_{0_\perp}^T x s^T - tx^T\Gamma_0\right).$$

and by similar arguments:

$$\nabla_{\text{vec}(B)} \exp\{ix^T\Gamma(B)s\}\Big|_{B=0} = \exp\{ix^T\Gamma_0^T s\} i\text{vec}\left(\Gamma_{0_\perp}^T x s^T\right).$$

Putting them together yields an expression for $\nabla_{\text{vec}(B)} D(s,t,0,P_0)$:

$$\text{vec}\Bigg\{\mathbb{E}_{P_0}\left[\exp\left(iX^T[\Gamma_0 s + \Gamma_{0_\perp} t]\right) i\left(\Gamma_{0_\perp}^T X s^T - tX^T\Gamma_0\right)\right] \\ - \exp\left(-\frac{1}{2}\|t\|_2^2\right)\mathbb{E}\left[\exp(iX^T\Gamma_0^T s) i\Gamma_{0_\perp}^T X s^T\right]\Bigg\}.$$

Notice the expression $\exp\left(iX^T\Gamma_0^T s\right) i\Gamma_{0_\perp}^T X s^T$ consists of $\Gamma_{0_\perp}^T X$ multiplying a function of $\Gamma_0^T X$. These quantities are independent: by assumption, P_0 has independent non-Gaussian and Gaussian components, with Γ_0 and $\Gamma_{0_\perp}$ spanning the true independent subspaces. Moreover, $\Gamma_{0_\perp}^T X \sim \mathcal{N}(0, I_{p-d})$. Thus, the expectation of this expression under P_0 is 0. Now calculate the expectation of the other term, using the independence of $\Gamma_0^T X$ and $\Gamma_{0_\perp}^T X$:

$$\mathbb{E}_{P_0}\left[\exp\left(iX^T[\Gamma_0 s + \Gamma_{0_\perp} t]\right) i\left(\Gamma_{0_\perp}^T X s^T - tX^T\Gamma_0\right)\right]$$

$$= \mathbb{E}_{P_0}\left[\exp(is^T\Gamma_0^T X)\right]\mathbb{E}_{P_0}\left[\exp(it^T\Gamma_{0_\perp}^T X) i\Gamma_{0_\perp}^T X\right] s^T$$

$$- t\mathbb{E}_{P_0}\left[\exp\left(it^T\Gamma_{0_\perp}^T X\right)\right]\mathbb{E}_{P_0}\left[\exp\left(is^T\Gamma_0^T X\right) iX^T\Gamma_0\right].$$

Since $\Gamma_{0_\perp}^T X \sim \mathcal{N}(0, I_{p-d})$, its characteristic function is:

$$\mathbb{E}_{P_0}\left[\exp\left(it^T\Gamma_{0_\perp}^T X\right)\right] = \exp\left(-\|t\|_2^2/2\right).$$

Furthermore, $\mathbb{E}_{P_0}\left[\exp\left(it^T\Gamma_{0_\perp}^T X\right) i\Gamma_{0_\perp}^T X\right] = \nabla_t \mathbb{E}_{P_0}\left[\exp\left(it^T\Gamma_{0_\perp}^T X\right)\right]$, from which we obtain:

$$\mathbb{E}_{P_0}\left[\exp\left(it^T\Gamma_{0_\perp}^T X\right) i\Gamma_{0_\perp}^T X\right] = -t\exp\{-\|t\|_2^2/2\}.$$

By assumption $\Gamma_0^T X \sim F$, hence:

$$\mathbb{E}_{P_0}\left[\exp\{is^T\Gamma_0^T X\}\right] = \mathcal{X}(s;F),$$

and:

$$\mathbb{E}_{P_0}\left[\exp\left(is^T\Gamma_0^T X\right) iX^T\Gamma_0\right] = \nabla\mathcal{X}(s;F)^T,$$

where $\nabla\mathcal{X}(s;F)$ is the gradient vector of $\mathcal{X}(s;F)$ with respect to s. Thus,

$$\begin{aligned}\mathbb{E}_{P_0}\Bigg[\exp\left(iX^T[\Gamma_0 s + \Gamma_{0_\perp} t]\right) i\left(\Gamma_{0_\perp}^T X s^T - tX^T\Gamma_0\right)\Bigg]& \\ = -t\exp\left(-\frac{1}{2}\|t\|_2^2\right)&\left[\mathcal{X}(s;F)s^T + \nabla\mathcal{X}(s;F)^T\right]\end{aligned}$$

Using the equation $\text{vec}(ABC) = \left(C^T \otimes A\right)\text{vec}(B)$ we obtain the expression:

$$\nabla_{\text{vec}(B)} D(s,t,0,P_0) = -\exp\left(-\frac{1}{2}\|t\|_2^2\right)\left[\left(\mathcal{X}(s;F)s + \nabla\mathcal{X}(s;F)\right)\otimes I_{p-d}\right]t.$$

Recall the original computation we needed to make:

$$\nabla^2_{\text{vec}(B)}\rho(\Gamma_0,P_0) = 2\iint \nabla_{\text{vec}(B)} D(s,t,0,P_0)\nabla_{\text{vec}(B)}\overline{D(s,t,0,P_0)}^T\phi_d(s)\phi_{p-d}(t)\text{d}s\text{d}t.$$

We can now substitute expressions for $\nabla_{\text{vec}(B)} D(s,t,0,P_0)$:

$$\begin{aligned}\nabla^2_{\text{vec}(B)}\rho(\Gamma_0,P_0) = 2\int &\phi_d(s)\left[\left(\mathcal{X}(s;F)s + \nabla\mathcal{X}(s;F)\right)\otimes I_{p-d}\right] \\ &\times\left\{\int \exp\{-\|t\|_2^2\}tt^T\phi_{p-d}(t)\text{d}t\right\}\overline{\left[\left(\mathcal{X}(s;F)s + \nabla\mathcal{X}(s;F)\right)^T\otimes I_{p-d}\right]}\text{d}s.\end{aligned}$$

The quantity $\exp\left(-\|t\|_2^2\right)\phi_{p-d}(t)$ is equal to $(1/3)^{(p-d)/2}$ times the density of the $\mathcal{N}(0,\frac{1}{3}I_{p-d})$ distribution. Hence:

$$\int \exp\{-\|t\|_2^2\} tt^T \phi_{p-d}(t)\mathrm{d}t = \left(\frac{1}{3}\right)^{\frac{p-d}{2}} \frac{1}{3} I_{p-d}.$$

The above result, in conjunction with the well-known identity $(A \otimes B)(C \otimes D) = (AC \otimes BD)$ (when the matrix dimensions conform properly–see [23]) yields:

$$\begin{aligned}
&\nabla^2_{\mathrm{vec}(B)}\rho(\Gamma_0, P_0)\\
&= \frac{2}{3}\left(\frac{1}{3}\right)^{\frac{p-d}{2}} \int \phi_d(s) \left[(\mathcal{X}(s;F)s + \nabla\mathcal{X}(s;F)) \overline{(\mathcal{X}(s;F)s + \nabla\mathcal{X}(s;F))}^T \otimes I_{p-d}\right] \mathrm{d}s\\
&= \frac{2}{3}\left(\frac{1}{3}\right)^{\frac{p-d}{2}} \left[\int \phi_d(s) \left(\mathcal{X}(s;F)s + \nabla\mathcal{X}(s;F)\right) \overline{(\mathcal{X}(s;F)s + \nabla\mathcal{X}(s;F))}^T \mathrm{d}s\right] \otimes I_{p-d}\\
&= \frac{2}{3}\left(\frac{1}{3}\right)^{\frac{p-d}{2}} M(F) \otimes I_{p-d},
\end{aligned}$$

where $M(F)$ is defined in the statement of Theorem 2.3.4. The end result of our calculations is the following asymptotic expansion of $\nabla^2_{\mathrm{vec}(B)}\rho(\Gamma_0, \widehat{P}_n)\mathrm{vec}(\widehat{B}_n)$:

$$\boxed{\begin{aligned}
\nabla^2_{\mathrm{vec}(B)}\rho(\Gamma_0, \widehat{P}_n)\mathrm{vec}(\widehat{B}_n) = \frac{2}{3}\left(\frac{1}{3}\right)^{\frac{p-d}{2}} &\left[M(F) \otimes I_{p-d}\right] \mathrm{vec}(\widehat{B}_n)\\
&+ O_P(n^{-1}\|\widehat{B}_n\|_F) + o_P(\|\widehat{B}_n\|_F).
\end{aligned}}$$

Show $\nabla_{\mathrm{vec}(B)}\rho(\Gamma_0, \widehat{P}_n) \approx \frac{1}{n}\sum_{i=1}^n \mathrm{vec}\left(\psi(X_i, \Gamma_0, P_0)\right)$.

Write:

$$\begin{aligned}
\nabla_{\mathrm{vec}(B)}\rho(\Gamma_0, \widehat{P}_n) &= \frac{1}{n^2}\sum_{i,j} \nabla_{\mathrm{vec}(B)} r(X_i, X_j, \Gamma_0)\\
&= \frac{1}{n^2}\sum_{i=1}^n \nabla_{\mathrm{vec}(B)} r(X_i, X_i, 0) + \frac{1}{n^2}\sum_{i \neq j} \nabla_{\mathrm{vec}(B)} r(X_i, X_j, \Gamma_0).
\end{aligned}$$

By Lemma B.3.2, for any value of B we have $\left|\frac{\partial}{\partial B_k} r(x, y, \Gamma(B))\right| \leq K(d,p)\left[\|x\|_2 + \|y\|_2\right]$ with the constant $K(d,p)$ depending only on d and p, and not on B or the index k. Therefore $\mathbb{E}_{P_0}\left[\|\nabla_{\mathrm{vec}(B)} r(X_i, X_j, \Gamma_0)\|_2\right] < \infty$ and the diagonal term $\frac{1}{n^2}\sum_{i=1}^n \nabla_{\mathrm{vec}(B)} r(X_i, X_i, \Gamma_0)$ is of order $O_P(n^{-1})$. So we focus on the off-diagonal terms. Notice that:

$$
\begin{aligned}
&\mathbb{E}_{P_0}\left[\nabla_{\mathrm{vec}(B)} r(X_1, X_2, 0)\right] = \iint \mathrm{d}P_0(x)\mathrm{d}P_0(y)\nabla_{\mathrm{vec}(B)} r(x, y, \Gamma_0)\\
&= \nabla_{\mathrm{vec}(B)}\rho(\Gamma(B), P_0)\Big|_{B=0}\\
&= \nabla_{\mathrm{vec}(B)}\left\{\iint D(s,t,B,P_0)\overline{D(s,t,B,P_0)}\phi_d(s)\phi_{p-d}(t)\mathrm{d}s\mathrm{d}t\right\}\Bigg|_{B=0}\\
&= 2\iint \nabla_{\mathrm{vec}(B)} D(s,t,B,P_0)\Big|_{B=0} D(s,t,0,P_0)\phi_d(s)\phi_{p-d}(t)\mathrm{d}s\mathrm{d}t\\
&= 0,
\end{aligned}
$$

since $D(s,t,0,P_0) = 0$ for all s and t. Therefore, the quantity $\frac{1}{n^2}\sum_{i\neq j}\nabla_{\mathrm{vec}(B)} r(X_i, X_j, \Gamma_0)$ has population mean 0. To facilitate the analysis, we symmetrize:

$$
\begin{aligned}
&\frac{1}{n^2}\sum_{i\neq j}\nabla_{\mathrm{vec}(B)} r(X_i, X_j, \Gamma_0)\\
&\quad= \frac{1}{n^2}\sum_{i<j}\nabla_{\mathrm{vec}(B)}\left(r(X_i, X_j, \Gamma_0) + r(X_j, X_i, \Gamma_0)\right).
\end{aligned}
$$

This is a vector-valued U-statistic with symmetric kernel $\nabla_{\mathrm{vec}(B)}\left(r(x, y, \Gamma_0) + r(y, x, \Gamma_0)\right)$. We established in Lemma B.3.2 that each component of this kernel is square integrable. Let $\psi(x, \Gamma(B), P) = \mathbb{E}_P\left[\nabla_B r(x, X, \Gamma(B))\right] + \mathbb{E}_P\left[\nabla_B r(X, x, \Gamma(B))\right]$ be defined as in the statement of Theorem 2.3.4. Then by Theorem 12.3, p. 162 in [40] we have, component wise,

$$
\frac{1}{n^2}\sum_{i<j}\frac{\partial}{\partial B_k}\left(r(X_i, X_j, \Gamma_0) + r(X_j, X_i, \Gamma_0)\right) = \frac{1}{n}\sum_{i=1}^{n}\psi_k(X_i, \Gamma_0, P_0) + o_P(n^{-\frac{1}{2}}).
$$

Since the dimension of the vectors $d(p-d)$ stays fixed, we can apply a union bound to obtain the asymptotic expansion for the whole vector:

$$
\boxed{\frac{1}{n^2}\sum_{i\neq j}\nabla_{\mathrm{vec}(B)}\, r(X_i, X_j, \Gamma_0) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}\left(\psi(X_i, \Gamma_0, P_0)\right) + o_P(n^{-\frac{1}{2}}).}
$$

Prove $M(F)$ is invertible.

We have obtained asymptotic representations for all the terms in the initial Taylor series. Putting them together yields:

$$
\begin{aligned}
0 = \frac{1}{n}\sum_{i=1}^{n} \operatorname{vec}\left(\psi(X_i, \Gamma_0, P_0)\right) + \frac{2}{3}\left(\frac{1}{3}\right)^{\frac{p-d}{2}} \left[M(F) \otimes I_{p-d}\right] \operatorname{vec}(\widehat{B}_n) \\
+ O_P(\|\widehat{B}_n\|_F^2) + o_P(\|\widehat{B}_n\|_F) + o_P(n^{-\frac{1}{2}}) + O_P(n^{-1}) + O_P(n^{-1}\|\widehat{B}_n\|_F).
\end{aligned}
$$

To obtain the asymptotic behavior of $\widehat{B}_n$ we need to invert $M(F) \otimes I_{p-d}$. If the inverse exists, it is equal to $M(F)^{-1} \otimes I_{p-d}$ using well-known properties of Kronecker products [23]. So we just have to show that $M(F)$ is invertible. Recall the definition:

$$
M(F) = \int \phi_d(s) \left(\mathcal{X}(s;F)s + \nabla\mathcal{X}(s;F)\right) \overline{\left(\mathcal{X}(s;F)s + \nabla\mathcal{X}(s;F)\right)}^T \mathrm{d}s.
$$

By inspection we see that $M(F)$ is symmetric and positive semidefinite. If we show it is strictly positive definite, we will have shown invertibility. Suppose $M(F)$ is *not* strictly positive definite. Then it has one zero eigenvalue. This implies there exists $w \in \mathbb{R}^d$ such that $w \neq 0$ and:

$$
\begin{aligned}
0 &= w^T M(F) w \\
&= \int \phi_d(s) \left[w^T \left(\mathcal{X}(s;F)s + \nabla\mathcal{X}(s;F)\right) \overline{\left(\mathcal{X}(s;F)s + \nabla\mathcal{X}(s;F)\right)}^T w\right] \mathrm{d}s \\
&= \int \phi_d(s) \left|w^T \left(\mathcal{X}(s;F)s + \nabla\mathcal{X}(s;F)\right)\right|^2 \mathrm{d}s.
\end{aligned}
$$

Thus $w^T\left(\mathcal{X}(s;F)s + \nabla\mathcal{X}(s;F)\right) = 0$ for all s except possibly on a set of measure zero under the $\mathcal{N}(0, I_d)$ distribution. However, since by assumption F has finite third moments, we know $\mathcal{X}(s;F)$ and $\nabla\mathcal{X}(s;F)$ are both continuous functions on $\mathbb{R}^d$. Therefore we must have $w^T\left(\mathcal{X}(s;F)s + \nabla\mathcal{X}(s;F)\right) = 0$ for all s.

Without loss of generality let $\|w\|_2 = 1$. Let the set of d-dimensional vectors $\{w_1, \ldots, w_{d-1}\}$ together with w form an orthonormal basis of $\mathbb{R}^d$; i.e. the matrix $\mathcal{W}$ formed by:

$$
\mathcal{W} = (w \; w_1 \; \ldots \; w_{d-1})^T
$$

is a $d \times d$ orthogonal matrix. We can write $\mathcal{X}(s;F)$ as:

$$
\begin{aligned}
\mathcal{X}(s;F) &= \mathbb{E}_F\left[\exp\{is^T V\}\right] \\
&= \mathbb{E}_F\left[\exp\{i(\mathcal{W}s)^T \mathcal{W}V\}\right] \\
&= \mathcal{X}(\mathcal{W}s; F_{\mathcal{W}}),
\end{aligned}
$$

where $F_{\mathcal{W}}$ is the distribution of $\mathcal{W}V$ when $V \sim F$. Let $\mathbf{z} = \mathcal{W}s$ and $z = w^T s$. Then $w^T s\mathcal{X}(s;F) = z\mathcal{X}(\mathbf{z}; F_{\mathcal{W}})$. Moreover, by the chain rule,

$$\frac{\partial}{\partial z}\mathcal{X}(\mathbf{z}, F_{\mathcal{W}}) = \frac{\partial}{\partial z}\mathcal{X}(s; F) = w^T \nabla \mathcal{X}(s; F).$$

Therefore:

$$0 = w^T \left(\mathcal{X}(s; F)s + \nabla \mathcal{X}(s; F) \right) = z\mathcal{X}(\mathbf{z}; F_{\mathcal{W}}) + \frac{\partial}{\partial z}\mathcal{X}(\mathbf{z}, F_{\mathcal{W}}).$$

This equation holds for all values of $\mathbf{z}$ and z. We now characterize what classes of distributions satisfy the above differential equation. Write the characteristic function of the transformed distribution $F_{\mathcal{W}}$ as:

$$\mathcal{X}(\mathbf{z}, F_{\mathcal{W}}) = e^{-z^2/2}\tilde{F}(\mathbf{z}).$$

Take the derivative in z:

$$\frac{\partial}{\partial z}\mathcal{X}(\mathbf{z}, F_{\mathcal{W}}) = -z\mathcal{X}(\mathbf{z}, F_{\mathcal{W}}) + e^{-z^2/2}\frac{\partial}{\partial z}\tilde{F}(\mathbf{z}).$$

Therefore we must have, since $e^{-z^2/2} > 0$ for all z,

$$\frac{\partial}{\partial z}\tilde{F}(\mathbf{z}) = 0.$$

So $\tilde{F}$ must be constant as a function of z. This implies $\mathcal{X}(\mathbf{z}, F_{\mathcal{W}})$ takes the form:

$$\mathcal{X}(\mathbf{z}, F_{\mathcal{W}}) = e^{-\frac{1}{2}z^2}\tilde{G}(z_1, \ldots, z_d).$$

The right hand side of the above display is the characteristic function of a a random vector with an independent $\mathcal{N}(0, 1)$ component. This is precisely the situation ruled out by the identifiability condition: no linear transformation of the non-Gaussian vector V should yield an independent component. Therefore, it must be that $M(F)$ is invertible.

Show $\widehat{B}_n \approx 3^{\frac{p-d}{2}} \frac{3}{2} \sum_{i=1}^n \psi(X_i, \Gamma_0, P_0)M(F)^{-1}$**.**

Recall the asymptotic representation:

$$0 = \frac{1}{n}\sum_{i=1}^n \operatorname{vec}\left(\psi(X_i, \Gamma_0, P_0)\right) + \frac{2}{3}\left(\frac{1}{3}\right)^{\frac{p-d}{2}} \left[M(F) \otimes I_{p-d}\right] \operatorname{vec}(\widehat{B}_n)$$
$$+ O_P(\|\widehat{B}_n\|_F^2) + o_P(\|\widehat{B}_n\|_F) + o_P(n^{-\frac{1}{2}}) + O_P(n^{-1}) + O_P(n^{-1}\|\widehat{B}_n\|_F).$$

Terms of stochastic order $O_P(n^{-1})$ are negligible relative to the leading terms. The consistency of the estimator (Theorem 2.3.2) implies $\|\widehat{B}_n\|_F \xrightarrow{P} 0$. Therefore $O_P(\|\widehat{B}_n\|_F^2) = o_P(\|\widehat{B}_n\|_F)$ and $O_P(n^{-1}\|\widehat{B}_n\|_F) = o_P(n^{-1})$; ignoring these negligible terms we can write:

$$\frac{2}{3}\left(\frac{1}{3}\right)^{\frac{p-d}{2}}\left[M(F)\otimes I_{p-d}\right]\mathrm{vec}(\widehat{B}_n)+o_P(\|\widehat{B}_n\|_F)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}\left(\psi(X_i,\Gamma_0,P_0)\right)+o_P(n^{-\frac{1}{2}}).$$

Multiply the above equation through by $3^{(p-d)/2}\frac{3}{2}\left[M(F)\otimes I_{p-d}\right]^{-1}$:

$$\begin{aligned}&\mathrm{vec}(\widehat{B}_n)+3^{(p-d)/2}\frac{3}{2}\left[M(F)\otimes I_{p-d}\right]^{-1}o_P(\|\widehat{B}_n\|_F)\\ =&3^{(p-d)/2}\frac{3}{2}\frac{1}{n}\sum_{i=1}^{n}\left[M(F)\otimes I_{p-d}\right]^{-1}\mathrm{vec}\left(\psi(X_i,\Gamma_0,P_0)\right)+\left[M(F)\otimes I_{p-d}\right]^{-1}o_P(n^{-\frac{1}{2}}).\end{aligned}$$

Since $\left[M(F)\otimes I_{p-d}\right]^{-1}$ is a deterministic matrix with finite entries whose dimension is fixed with n, the term $\left[M(F)\otimes I_{p-d}\right]^{-1}o_P(n^{-\frac{1}{2}})$ is still $o_P(n^{-\frac{1}{2}})$. By the same reasoning, $\left[M(F)\otimes I_{p-d}\right]^{-1}o_P(\|\widehat{B}_n\|_F)=o_P(\|\widehat{B}_n\|_F)$. Therefore, since:

$$\left\|\mathrm{vec}(\widehat{B}_n)+o_P(\|\widehat{B}_n\|_F)\right\|_2=\|\widehat{B}_n\|_F\left|1+o_P(1)\right|;$$

and since:

$$\begin{aligned}&\left\|3^{(p-d)/2}\frac{3}{2}\frac{1}{n}\sum_{i=1}^{n}\left[M(F)\otimes I_{p-d}\right]^{-1}\mathrm{vec}\left(\psi(X_i,\Gamma_0,P_0)\right)+o_P(n^{-\frac{1}{2}})\right\|_2\\ =&\left\|3^{(p-d)/2}\frac{3}{2}\frac{1}{n}\sum_{i=1}^{n}\left[M(F)\otimes I_{p-d}\right]^{-1}\mathrm{vec}\left(\psi(X_i,\Gamma_0,P_0)\right)\right\|_2\left|1+o_P(1)\right|\end{aligned}$$

($\sum_{i=1}^{n}\left[M(F)\otimes I_{p-d}\right]^{-1}\mathrm{vec}\left(\psi(X_i,\Gamma_0,P_0)\right)=O_P(n^{-\frac{1}{2}})$ by the central limit theorem), it follows that $\|\widehat{B}_n\|_F=O_P(n^{-\frac{1}{2}})$. We conclude:

$$\mathrm{vec}(\widehat{B}_n)=3^{\frac{p-d}{2}}\frac{3}{2}\frac{1}{n}\sum_{i=1}^{n}\left[M(F)^{-1}\otimes I_{p-d}\right]\mathrm{vec}\left(\psi(X_i,\Gamma_0,P_0)\right)+o_P(n^{-\frac{1}{2}}).$$

The above expression is in the vectorized form. We can use the identity $\mathrm{vec}(ABC)=(C^T\otimes A)\mathrm{vec}(B)$ to obtain the asymptotic expansion of $\widehat{B}_n$, thereby completing the proof:

$$\widehat{B}_n=3^{\frac{p-d}{2}}\frac{3}{2}\sum_{i=1}^{n}\psi(X_i,\Gamma_0,P_0)M(F)^{-1}+o_P(n^{-\frac{1}{2}}).$$

□

We now prove Theorem 2.3.8, which gave the asymptotic expansion of CHFNGCA when the population mean and covariance are unknown.

Proof of Theorem 2.3.8. Recall $\widehat{X}_i = \widehat{\Sigma}^{-1/2}(X_i - \widehat{\mu})$ and $\tilde{X}_i = \Sigma_0^{-1/2}(X_i - \mu_0)$. At $\widehat{\tilde{B}}_n$:

$$\begin{aligned}
0 &= \nabla_{\text{vec}(B)} \rho(\widehat{\tilde{\Gamma}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})) \\
&= \frac{1}{n^2} \sum_{i,j} \nabla_{\text{vec}(B)} r(\widehat{X}_i, \widehat{X}_j, \widehat{\tilde{\Gamma}}_n)
\end{aligned}$$

We expand $\widehat{X}_i$ and $\widehat{X}_j$ around $\tilde{X}_i$ and $\tilde{X}_j$ to work with the population mean and population covariance. Let $\nabla_x r(x, y, \Gamma)$ be the gradient vector of r with respect to the first argument, and $\nabla_y r(x, y, \Gamma)$ be the gradient vector of r with respect to the second argument. Then we have:

$$\begin{aligned}
0 &= \frac{1}{n^2} \sum_{i,j} \nabla_{\text{vec}(B)} r(\tilde{X}_i, \tilde{X}_j, \widehat{\tilde{\Gamma}}_n) \\
&+ \frac{1}{n^2} \sum_{i,j} \left[\nabla_{\text{vec}(B)} \nabla_x^T r(\tilde{X}_i, \tilde{X}_j, \widehat{\tilde{\Gamma}}_n) \right] (\widehat{X}_i - \tilde{X}_i) \\
&+ \frac{1}{n^2} \sum_{i,j} \left[\nabla_{\text{vec}(B)} \nabla_y^T r(\tilde{X}_i, \tilde{X}_j, \widehat{\tilde{\Gamma}}_n) \right] (\widehat{X}_j - \tilde{X}_j) \\
&+ R_1(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})),
\end{aligned}$$

where $R_1(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu}))$ is a remainder term whose kth component can be expressed in integral form as:

$$\begin{aligned}
&\left(R_1(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})) \right)_k = \\
&\frac{1}{n^2} \sum_{i,j} \left\{ \left(\widehat{X}_i - \tilde{X}_i, \widehat{X}_j - \tilde{X}_j \right)^T \left[\int_0^1 \nabla^2_{(x,y)} \frac{\partial}{\partial B_k} r(\widehat{X}_i^\xi, \widehat{X}_j^\xi, \widehat{\tilde{\Gamma}}_n) \mathrm{d}\xi \right] \left(\widehat{X}_i - \tilde{X}_i, \widehat{X}_j - \tilde{X}_j \right) \right\}.
\end{aligned}$$

The operator $\nabla^2_{(x,y)}$ takes the Hessian matrix in all the x, y coordinates, while $\widehat{X}_i^\xi = \xi \widehat{X}_i + (1 - \xi) \tilde{X}_i$. Take another Taylor expansion of $\widehat{\tilde{\Gamma}}_n = \tilde{\Gamma}(\widehat{\tilde{B}}_n)$ about 0 to work with the true non-Gaussian subspace:

$$
\begin{aligned}
&\frac{1}{n^2}\sum_{i,j}\left[\nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_i,\tilde{X}_j,\widehat{\tilde{\Gamma}}_n)\right](\widehat{X}_i-\tilde{X}_i)\\
&+\frac{1}{n^2}\sum_{i,j}\left[\nabla_{\mathrm{vec}(B)}\nabla_y^T r(\tilde{X}_i,\tilde{X}_j,\widehat{\tilde{\Gamma}}_n)\right](\widehat{X}_j-\tilde{X}_j)\\
&=\frac{1}{n^2}\sum_{i,j}\left[\nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_i,\tilde{X}_j,\tilde{\Gamma}_0)\right](\widehat{X}_i-\tilde{X}_i)\\
&+\frac{1}{n^2}\sum_{i,j}\left[\nabla_{\mathrm{vec}(B)}\nabla_y^T r(\tilde{X}_i,\tilde{X}_j,\tilde{\Gamma}_0)\right](\widehat{X}_j-\tilde{X}_j)\\
&+R_2(\widehat{\tilde{B}}_n,\widehat{P}_n(\widehat{\Sigma},\widehat{\mu})),
\end{aligned}
$$

where $R_2(\widehat{\tilde{B}}_n,\widehat{P}_n(\widehat{\Sigma},\widehat{\mu}))$ is a remainder term whose kth component is equal to:

$$
\frac{1}{n^2}\sum_{i,j}\left\{\widehat{\tilde{B}}_n^T\left[\int_0^1\nabla_{\mathrm{vec}(B)}\nabla_{(x,y)}^T\frac{\partial}{\partial B_k}r\left(\widehat{X}_i,\widehat{X}_j,\tilde{\Gamma}\left(\xi\widehat{\tilde{B}}_n\right)\right)\mathrm{d}\xi\right](\widehat{X}_i-\tilde{X}_i,\widehat{X}_j-\tilde{X}_j)\right\}.
$$

At this point, expand $\widehat{X}_i-\tilde{X}_i$:

$$
\begin{aligned}
\widehat{X}_i-\tilde{X}_i&=\left(\widehat{\Sigma}^{-1/2}-\Sigma_0^{-1/2}\right)(X_i-\mu_0)+\Sigma_0^{-1/2}(\mu_0-\hat{\mu})+\left(\widehat{\Sigma}^{-1/2}-\Sigma_0^{-1/2}\right)(\mu_0-\hat{\mu})\\
&=\left(\widehat{\Sigma}^{-1/2}-\Sigma_0^{-1/2}\right)\Sigma_0^{1/2}\tilde{X}_i+\Sigma_0^{-1/2}(\mu_0-\hat{\mu})+\left(\widehat{\Sigma}^{-1/2}-\Sigma_0^{-1/2}\right)(\mu_0-\hat{\mu}).
\end{aligned}
$$

We now write out the full Taylor expansion as follows:

$$
\begin{aligned}
0 &= \frac{1}{n^2}\sum_{i,j}\nabla_{\text{vec}(B)} r(\tilde{X}_i, \tilde{X}_j, \widehat{\tilde{\Gamma}}_n) \\
&+ \frac{1}{n^2}\sum_{i,j}\left[\nabla_{\text{vec}(B)}\nabla_x^T r(\tilde{X}_i, \tilde{X}_j, \tilde{\Gamma}_0)\left(\widehat{\Sigma}^{-1/2} - \Sigma_0^{-1/2}\right)\Sigma_0^{1/2}\tilde{X}_i\right] \\
&+ \frac{1}{n^2}\sum_{i,j}\left[\nabla_{\text{vec}(B)}\nabla_y^T r(\tilde{X}_j, \tilde{X}_j, \tilde{\Gamma}_0)\left(\widehat{\Sigma}^{-1/2} - \Sigma_0^{-1/2}\right)\Sigma_0^{1/2}\tilde{X}_j\right] \\
&+ \frac{1}{n^2}\sum_{i,j}\left[\nabla_{\text{vec}(B)}\nabla_x^T r(\tilde{X}_i, \tilde{X}_j, \tilde{\Gamma}_0)\right]\Sigma_0^{-1/2}(\mu_0 - \hat{\mu}) \\
&+ \frac{1}{n^2}\sum_{i,j}\left[\nabla_{\text{vec}(B)}\nabla_y^T r(\tilde{X}_j, \tilde{X}_j, \tilde{\Gamma}_0)\right]\Sigma_0^{-1/2}(\mu_0 - \hat{\mu}) \\
&+ \frac{1}{n^2}\sum_{i,j}\left[\nabla_{\text{vec}(B)}\nabla_x^T r(\tilde{X}_i, \tilde{X}_j, \tilde{\Gamma}_0)\right]\left(\widehat{\Sigma}^{-1/2} - \Sigma_0^{-1/2}\right)(\mu_0 - \hat{\mu}) \\
&+ \frac{1}{n^2}\sum_{i,j}\left[\nabla_{\text{vec}(B)}\nabla_y^T r(\tilde{X}_j, \tilde{X}_j, \tilde{\Gamma}_0)\right]\left(\widehat{\Sigma}^{-1/2} - \Sigma_0^{-1/2}\right)(\mu_0 - \hat{\mu}) \\
&+ R_1(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})) + R_2(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu}))
\end{aligned}
$$

We deal with the terms separately:

Show $R_2(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu}))$ is negligible.

Recall:

$$
\begin{aligned}
&\left(R_2(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu}))\right)_k = \\
&\qquad \frac{1}{n^2}\sum_{i,j}\left\{\widehat{\tilde{B}}_n^T\left[\int_0^1 \nabla_{\text{vec}(B)}\nabla_{(x,y)}^T \frac{\partial}{\partial B_k} r\left(\widehat{X}_i, \widehat{X}_j, \tilde{\Gamma}\left(\xi\widehat{\tilde{B}}_n\right)\right)\mathrm{d}\xi\right](\widehat{X}_i - \tilde{X}_i, \widehat{X}_j - \tilde{X}_j)\right\},
\end{aligned}
$$

where:

$$
\begin{aligned}
\widehat{X}_i - \tilde{X}_i &= \left(\widehat{\Sigma}^{-1/2} - \Sigma_0^{-1/2}\right)(X_i - \mu_0) + \Sigma_0^{-1/2}(\mu_0 - \hat{\mu}) + \left(\widehat{\Sigma}^{-1/2} - \Sigma_0^{-1/2}\right)(\mu_0 - \hat{\mu}) \\
&= \left(\widehat{\Sigma}^{-1/2} - \Sigma_0^{-1/2}\right)\Sigma_0^{1/2}\tilde{X}_i + \Sigma_0^{-1/2}(\mu_0 - \hat{\mu}) + \left(\widehat{\Sigma}^{-1/2} - \Sigma_0^{-1/2}\right)(\mu_0 - \hat{\mu}).
\end{aligned}
$$

By Lemma B.3.3 each entry of the matrix $\nabla_{\text{vec}(B)}\nabla_{(x,y)}^T \frac{\partial}{\partial B_k} r(\widehat{X}_i, \widehat{X}_j, \tilde{\Gamma}(\xi\widehat{\tilde{B}}_n))$ is bounded in absolute value by $K(p,d)\left(\|\widehat{X}_i\|_2^2 + \|\widehat{X}_j\|_2^2\right)$ where $K(p,d)$ is some constant that depends only on p and d. Therefore, for each index k:

$$
\begin{aligned}
&\left| \left(R_2(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})) \right)_k \right| \\
&\leq \frac{1}{n^2} \sum_{i,j} \left\{ \left\| \int_0^1 \nabla_{\mathrm{vec}(B)} \nabla^T_{(x,y)} \frac{\partial}{\partial B_k} r(\widehat{X}_i, \widehat{X}_j, \tilde{\Gamma}(\xi \widehat{\tilde{B}}_n)) \mathrm{d}\xi \right\|_F \left\| \left(\widehat{X}_i - \tilde{X}_i, \widehat{X}_j - \tilde{X}_j \right) \right\|_2 \right\} \|\widehat{\tilde{B}}_n\|_F \\
&\leq K'(p,d) \Bigg\{ \left(\frac{1}{n} \sum_{i=1}^n \|\widehat{X}_i\|_2^2 \|\tilde{X}_i\|_2 \right) \|\widehat{\Sigma}^{-1/2} - \Sigma_0^{-1/2}\|_2 \|\Sigma_0^{1/2}\|_F \\
&+ \left(\frac{1}{n} \sum_{i=1}^n \|\widehat{X}_i\|_2^2 \right) \|\Sigma_0^{1/2}\|_2 \|\hat{\mu} - \mu_0\|_2 + \left(\frac{1}{n} \sum_{i=1}^n \|\widehat{X}_i\|_2^2 \right) \|\widehat{\Sigma}^{-1/2} - \Sigma_0^{-1/2}\|_2 \|\hat{\mu} - \mu_0\|_2 \Bigg\} \|\widehat{\tilde{B}}_n\|_F.
\end{aligned}
$$

We systematically investigate the size of each of the above terms. By Holder's inequality, $\frac{1}{n}\sum_{i=1}^n \|\widehat{X}_i\|_2^2 \|\tilde{X}_i\|_2 \leq \left(\frac{1}{n}\sum_{i=1}^n \|\widehat{X}_i\|_2^3 \right)^{2/3} \left(\frac{1}{n}\sum_{i=1}^n \|\tilde{X}_i\|_2^3 \right)^{1/3}$, and by the law of large numbers:

$$
\frac{1}{n} \sum_{i=1}^n \|\tilde{X}_i\|_2^3 \xrightarrow{P} \mathbb{E}_{\tilde{P}_0} \left[\|\tilde{X}\|_2^3 \right] < \infty
$$

From the construction of $\widehat{X}_i$ we have,

$$
\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \|\widehat{X}_i\|_2^3 &\leq \|\widehat{\Sigma}^{-1/2}\|_2^3 \frac{1}{n} \sum_{i=1}^n \|X_i - \hat{\mu}\|_2^3 \\
&\leq \|\widehat{\Sigma}^{-1/2}\|_2^3 \frac{1}{n} \sum_{i=1}^n \left(\|X_i - \mu_0\|_2 + \|\hat{\mu} - \mu_0\|_2 \right)^3 \\
&= \|\widehat{\Sigma}^{-1/2}\|_2^3 \frac{1}{n} \sum_{i=1}^n \Bigg(\|X_i - \mu_0\|_2^3 + \|\hat{\mu} - \mu_0\|_2^3 \\
&\quad + 3\|X_i - \mu_0\|_2^2 \|\widehat{\mu} - \mu_0\|_2 + 3\|X_i - \mu_0\|_2 \|\widehat{\mu} - \mu_0\|_2^2 \Bigg).
\end{aligned}
$$

Since $\widehat{\Sigma}$ is consistent and Σ_0 is positive definite, we have $\|\widehat{\Sigma}^{-1/2}\|_2 \xrightarrow{P} \|\Sigma_0^{-1/2}\|_2$. Also:

$$
\frac{1}{n} \sum_{i=1}^n \|X_i - \mu_0\|_2^3 \xrightarrow{P} \mathbb{E}_{P_0} \left[\|X - \mu_0\|_2^3 \right] < \infty.
$$

The other terms go to zero, since $\|\widehat{\mu} - \mu\|_2 = o_P(1)$ (more precisely, $\|\widehat{\mu} - \mu\|_2 = O_P(n^{-1/2})$).

We have thus far shown $\frac{1}{n}\sum_{i=1}^n \|\widehat{X}_i\|_2^2 \|\tilde{X}_i\|_2 = O_P(1)$ and $\frac{1}{n}\sum_{i=1}^n \|\widehat{X}_i\|_2^2 = O_P(1)$. Since we assume the data have finite fourth moments, the sample covariance $\widehat{\Sigma}$ is $\sqrt{n}$-consistent. Therefore,

$$\|\widehat{\Sigma}^{-1/2} - \Sigma_0^{-1/2}\|_2 = O_P(n^{1/2}).$$

$\|\hat{\mu} - \mu_0\|_2 = O_P(n^{1/2})$ implies the product term $\|\widehat{\Sigma}^{-1/2} - \Sigma_0^{-1/2}\|_2 \|\hat{\mu} - \mu_0\|_2 = O_P(n^{-1})$. Finally, by Theorem 2.3.7, $\|\widehat{\tilde{B}}_n\|_F \xrightarrow{P} 0$. Therefore, the leading term of the remainder kth coordinate of R_2 is order $o_P(n^{-1/2})$:

$$R_2(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu}))_k = o_P(n^{-1/2}).$$

Since the dimension of the vector $R_2(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu}))$ is fixed as $n \to \infty$, by a union bound we obtain the stochastic order of the whole vector:

$$\boxed{R_2(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})) = o_P(n^{-1/2}).}$$

Show $R_1(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu}))$ is negligible.

The kth entry of $R_1(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu}))$ has the form:

$$\begin{aligned}
&\left(R_1(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu}))\right)_k = \\
&\frac{1}{n^2} \sum_{i,j} \left\{ \left(\widehat{X}_i - \tilde{X}_i, \widehat{X}_j - \tilde{X}_j\right)^T \left[\int_0^1 \nabla^2_{(x,y)} \frac{\partial}{\partial B_k} r(\widehat{X}_i^\xi, \widehat{X}_j^\xi, \widehat{\tilde{\Gamma}}_n) \mathrm{d}\xi\right] \left(\widehat{X}_i - \tilde{X}_i, \widehat{X}_j - \tilde{X}_j\right) \right\};
\end{aligned}$$

By Lemma B.3.3 each entry of the matrix $\int_0^1 \nabla^2_{(x,y)} \frac{\partial}{\partial B_k} r(\widehat{X}_i^\xi, \widehat{X}_j^\xi, \widehat{\tilde{\Gamma}}_n) \mathrm{d}\xi$ has the upper bound

$$\begin{aligned}
&\int_0^1 \nabla^2_{(x,y)} \frac{\partial}{\partial B_k} r(\widehat{X}_i^\xi, \widehat{X}_j^\xi, \widehat{\tilde{\Gamma}}_n) \mathrm{d}\xi \\
&\leq K_1(p,d) \int_0^1 \left(\|\widehat{X}_i^\xi\|_2 + \|\widehat{X}_j^\xi\|_2\right) \mathrm{d}\xi \\
&\leq K_1(p,d) \int_0^1 \left[\left(\xi\|\widehat{X}_i\|_2 + (1-\xi)\|\tilde{X}_i\|_2\right) + \left(\xi\|\widehat{X}_j\|_2 + (1-\xi)\|\tilde{X}_j\|_2\right)\right] \mathrm{d}\xi \\
&\leq K_1(p,d)/2 \left[\left(\|\widehat{X}_i\|_2 + \|\tilde{X}_i\|_2\right) + \left(\|\widehat{X}_j\|_2 + \|\tilde{X}_j\|_2\right)\right].
\end{aligned}$$

Hence:

$$\left|\left(R_1(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu}))\right)_k\right| \leq K_1'(p,d)\left[\frac{1}{n}\sum_{i=1}^{n}\left(\|\widehat{X}_i\|_2 + \|\tilde{X}_i\|_2\right)\right]\|\left(\widehat{X}_i - \tilde{X}_i, \widehat{X}_j - \tilde{X}_j\right)\|_2^2.$$

We have shown that $\frac{1}{n}\sum_{i=1}^{n}\|\widehat{X}_i\|_2 = O_P(1)$ and $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{X}_i\|_2 = O_P(1)$. We have also shown the leading order term of $\widehat{X}_i - \tilde{X}_i$ to be order $O_P(n^{-1/2})$. Thus, $\left|\left(R_1(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu}))\right)_k\right| = O_P(n^{-1})$, and by a union bound,

$$\boxed{R_1(\widehat{\tilde{B}}_n, \widehat{P}_n(\widehat{\Sigma}, \widehat{\mu})) = O_P(n^{-1}).}$$

Show

$$\frac{1}{n^2}\sum_{i,j}\left[\nabla_{\text{vec}(B)}\nabla_x^T r(\tilde{X}_i, \tilde{X}_j, \tilde{\Gamma}_0) + \nabla_{\text{vec}(B)}\nabla_y^T r(\tilde{X}_i, \tilde{X}_j, \tilde{\Gamma}_0)\right]\left(\widehat{\Sigma}^{-1/2} + \Sigma_0^{-1/2}\right)(\mu_0 - \hat{\mu})$$

is negligible.

The matrix

$$\frac{1}{n^2}\sum_{i,j}\left[\nabla_{\text{vec}(B)}\nabla_x^T r(\tilde{X}_i, \tilde{X}_j, \tilde{\Gamma}_0) + \nabla_{\text{vec}(B)}\nabla_y^T r(\tilde{X}_i, \tilde{X}_j, \tilde{\Gamma}_0)\right]$$

is a V-statistic with integrable entries (see Lemma B.3.3). The $i = j$ diagonal terms are $O_P(n^{-1})$ while the off diagonal terms,

$$\frac{1}{n^2}\sum_{i\neq j}\left[\nabla_{\text{vec}(B)}\nabla_x^T r(\tilde{X}_i, \tilde{X}_j, \tilde{\Gamma}_0) + \nabla_{\text{vec}(B)}\nabla_y^T r(\tilde{X}_i, \tilde{X}_j, \tilde{\Gamma}_0)\right]$$

are $O_P(1)$ by the law or large numbers for U-statistics [24]. But, the product $\left(\widehat{\Sigma}^{-1/2} - \Sigma_0^{-1/2}\right)(\mu_0 - \hat{\mu})$ is $O_P(n^{-1})$, which shows:

$$\boxed{\begin{aligned}&\frac{1}{n^2}\sum_{i,j}\left[\nabla_{\text{vec}(B)}\left(\nabla_x^T r(\tilde{X}_i, \tilde{X}_j, \tilde{\Gamma}_0) + \nabla_y^T r(\tilde{X}_i, \tilde{X}_j, \tilde{\Gamma}_0)\right)\right]\left(\widehat{\Sigma}^{-1/2} - \Sigma_0^{-1/2}\right)(\mu_0 - \hat{\mu})\\ &\qquad = O_P(n^{-1}).\end{aligned}}$$

Show

$$
\begin{aligned}
&\frac{1}{n^2}\sum_{i,j}\left\{\left[\nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_i,\tilde{X}_j,\tilde{\Gamma}_0)\right]\left(\widehat{\Sigma}^{-1/2}-\Sigma_0^{-1/2}\right)\Sigma_0^{1/2}\tilde{X}_i\right\}\\
+&\frac{1}{n^2}\sum_{i,j}\left\{\left[\nabla_{\mathrm{vec}(B)}\nabla_y^T r(\tilde{X}_i,\tilde{X}_j,\tilde{\Gamma}_0)\right]\left(\widehat{\Sigma}^{-1/2}-\Sigma_0^{-1/2}\right)\Sigma_0^{1/2}\tilde{X}_j\right\}\\
\approx&-\frac{1}{2}\mathbb{E}_{\tilde{P}_0}\left[(\tilde{X})^T\otimes\nabla_x^T\mathrm{vec}\left(\psi(\tilde{X},\tilde{\Gamma}_0,\tilde{P}_0)\right)\right]\left(\Sigma_0^{\frac{1}{4}}\otimes\Sigma_0^{-\frac{1}{4}}\right)\mathrm{vec}\left(\frac{1}{n}\sum_{i=1}^n\tilde{X}_i\tilde{X}_i^T-I_p\right).
\end{aligned}
$$

We have a simple identity for the sample covariance matrix:

$$
\begin{aligned}
\widehat{\Sigma}&=\frac{1}{n}\sum_{i=1}^n(X_i-\hat{\mu})(X_i-\hat{\mu})^T\\
&=\Sigma_0^{1/2}\left(\frac{1}{n}\sum_{i=1}^n(\tilde{X}_i-\widehat{\tilde{\mu}})(\tilde{X}_i-\widehat{\tilde{\mu}})^T\right)\Sigma_0^{1/2}\\
&=\Sigma_0^{1/2}\widehat{\tilde{\Sigma}}\Sigma_0^{1/2},
\end{aligned}
$$

where $\widehat{\tilde{\mu}}=\frac{1}{n}\sum_{i=1}^n\tilde{X}_i$ is the sample mean, and $\widehat{\tilde{\Sigma}}$ the sample covariance matrix, on the *whitened* data. Therefore, we can write:

$$
\begin{aligned}
&\frac{1}{n^2}\sum_{i,j}\left\{\left[\nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_i,\tilde{X}_j,\tilde{\Gamma}_0)\right]\left(\widehat{\Sigma}^{-\frac{1}{2}}-\Sigma_0^{-\frac{1}{2}}\right)\Sigma_0^{\frac{1}{2}}\tilde{X}_i\right\}\\
&\qquad=\frac{1}{n^2}\sum_{i,j}\left\{\left[\nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_i,\tilde{X}_j,\tilde{\Gamma}_0)\right]\Sigma_0^{-\frac{1}{4}}\left(\widehat{\tilde{\Sigma}}^{-\frac{1}{2}}-I_p\right)\Sigma_0^{\frac{1}{4}}\tilde{X}_i\right\}.
\end{aligned}
$$

Using the formula $\mathrm{vec}(ABC)=(C^T\otimes A)\mathrm{vec}(B)$ twice, we can rewrite the above as:

$$
\begin{aligned}
&\frac{1}{n^2}\sum_{i,j}\left\{\left[\nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_i,\tilde{X}_j,\tilde{\Gamma}_0)\right]\Sigma_0^{-\frac{1}{4}}\left(\widehat{\tilde{\Sigma}}^{-\frac{1}{2}}-I_p\right)\Sigma_0^{\frac{1}{4}}\tilde{X}_i\right\}\\
&\qquad=\frac{1}{n^2}\sum_{i,j}\left[\tilde{X}_i^T\otimes\nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_i,\tilde{X}_j,\tilde{\Gamma}_0)\right]\left(\Sigma_0^{\frac{1}{4}}\otimes\Sigma_0^{-\frac{1}{4}}\right)\mathrm{vec}\left(\widehat{\tilde{\Sigma}}^{-\frac{1}{2}}-I_p\right).
\end{aligned}
$$

The matrix:

$$\frac{1}{n^2}\sum_{i,j}\left[\tilde{X}_i^T \otimes \nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_i, \tilde{X}_j, \tilde{\Gamma}_0)\right]$$

is a V-statistic (not necessarily symmetric). By Lemma B.3.3 we see that each entry can be bounded by $\frac{1}{n}\sum_{i=1}^n \|\tilde{X}_i\|_2^2$ up to a constant that only depends on p and d. Therefore, the $i = j$ terms have order $O_P(n^{-1})$. Since $\hat{\tilde{\Sigma}} - I_p = O_P(n^{-1/2})$, we can effectively ignore these diagonal terms. For $i \neq j$ the matrix converges to its expectation by the law of large numbers for U-statistics. The error from replacing the U-statistic term with its expectation is $o_P(1)$, and $\hat{\tilde{\Sigma}} - I_p = O_P(n^{-1/2})$, which implies:

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i,j}\left[\tilde{X}_i^T \otimes \nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_i, \tilde{X}_j, \tilde{\Gamma}_0)\right]\left(\Sigma_0^{\frac{1}{4}} \otimes \Sigma_0^{-\frac{1}{4}}\right)\mathrm{vec}\left(\hat{\tilde{\Sigma}}^{-\frac{1}{2}} - I_p\right)\\
&= \mathbb{E}_{\tilde{P}_0\times\tilde{P}_0}\left[\tilde{X}_1^T \otimes \nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_1, \tilde{X}_2, \tilde{\Gamma}_0)\right]\left(\Sigma_0^{\frac{1}{4}} \otimes \Sigma_0^{-\frac{1}{4}}\right)\mathrm{vec}\left(\hat{\tilde{\Sigma}}^{-\frac{1}{2}} - I_p\right) + o_P(n^{-1/2}).
\end{aligned}$$

The other term can be represented the same way via the same arguments:

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i,j}\left[\tilde{X}_j^T \otimes \nabla_{\mathrm{vec}(B)}\nabla_y^T r(\tilde{X}_i, \tilde{X}_j, \tilde{\Gamma}_0)\right]\left(\Sigma_0^{\frac{1}{4}} \otimes \Sigma_0^{-\frac{1}{4}}\right)\mathrm{vec}\left(\hat{\tilde{\Sigma}}^{-\frac{1}{2}} - I_p\right)\\
&= \mathbb{E}_{\tilde{P}_0\times\tilde{P}_0}\left[\tilde{X}_2^T \otimes \nabla_{\mathrm{vec}(B)}\nabla_y^T r(\tilde{X}_1, \tilde{X}_2, \tilde{\Gamma}_0)\right]\left(\Sigma_0^{\frac{1}{4}} \otimes \Sigma_0^{-\frac{1}{4}}\right)\mathrm{vec}\left(\hat{\tilde{\Sigma}}^{-\frac{1}{2}} - I_p\right) + o_P(n^{-1/2}).
\end{aligned}$$

We draw a connection between the matrices $\mathbb{E}_{\tilde{P}_0}\left[\tilde{X}_1^T \otimes \nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_1, \tilde{X}_2, \tilde{\Gamma}_0)\right]$ and $\mathbb{E}_{\tilde{P}_0}\left[\tilde{X}_2^T \otimes \nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_1, \tilde{X}_2, \tilde{\Gamma}_0)\right]$ and the ψ function. Recall the definition of $\psi(x, \tilde{\Gamma}(B), P)$:

$$\psi(x, \tilde{\Gamma}(B), P) = \mathbb{E}_P\left[\nabla_B r(x, X, \tilde{\Gamma}(B))\right] + \mathbb{E}_P\left[\nabla_B r(X, x, \Gamma(\tilde{B}))\right]$$

We compute the partial derivatives of $\psi(x, \tilde{\Gamma}_0, \tilde{P}_0)$ in x. Since $\nabla_B r(x, y, \tilde{\Gamma}(B))$ is dominated (up to constants) by the $\tilde{P}_0$-integrable function $\|x\|_2 + \|y\|_2$ we can use the Dominated Convergence Theorem to exchange the derivative and expectation operators and obtain the expression:

$$\nabla_x^T \mathrm{vec}(\psi(x, \tilde{\Gamma}_0, \tilde{P}_0)) = \mathbb{E}_{\tilde{P}_0}\left[\nabla_{\mathrm{vec}(B)}\nabla_x^T r(x, X, \tilde{\Gamma}(B))\right] + \mathbb{E}_{\tilde{P}_0}\left[\nabla_{\mathrm{vec}(B)}\nabla_y^T r(X, x, \Gamma(\tilde{B}))\right].$$

Therefore:

$$
\begin{aligned}
&\mathbb{E}_{\tilde{P}_0\times\tilde{P}_0}\left[\tilde{X}_1^T \otimes \nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_1,\tilde{X}_2,\tilde{\Gamma}_0)\right] + \mathbb{E}_{\tilde{P}_0}\left[\tilde{X}_2^T \otimes \nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_1,\tilde{X}_2,\tilde{\Gamma}_0)\right]\\
&\mathbb{E}_{\tilde{P}_0\times\tilde{P}_0}\left[\tilde{X}_1^T \otimes \nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_1,\tilde{X}_2,\tilde{\Gamma}_0)\right] + \mathbb{E}_{\tilde{P}_0}\left[\tilde{X}_1^T \otimes \nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_2,\tilde{X}_1,\tilde{\Gamma}_0)\right]\\
&\quad= \mathbb{E}_{\tilde{P}_0}\left[\tilde{X}_1 \otimes \left\{\mathbb{E}\left[\nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_1,\tilde{X}_2,\tilde{\Gamma}_0)\middle|\tilde{X}_1\right] + \mathbb{E}\left[\nabla_{\mathrm{vec}(B)}\nabla_y^T r(\tilde{X}_2,\tilde{X}_1,\tilde{\Gamma}_0)\middle|\tilde{X}_1\right]\right\}\right]\\
&\quad= \mathbb{E}_{\tilde{P}_0}\left[(\tilde{X}_1)^T \otimes \nabla_x^T \mathrm{vec}\left(\psi(\tilde{X}_1,\tilde{\Gamma}_0,\tilde{P}_0)\right)\right].
\end{aligned}
$$

We return to the quantity $\widehat{\tilde{\Sigma}}^{-1/2} - I_p$. To obtain its asymptotic behavior, begin by looking at the normalized quantity $\sqrt{n}\mathrm{vec}\left(\widehat{\tilde{\Sigma}} - I_p\right)$. It is asymptotically normal: to see this, write:

$$
\begin{aligned}
\sqrt{n}\mathrm{vec}\left(\widehat{\tilde{\Sigma}} - I_p\right) &= \sqrt{n}\ \mathrm{vec}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{X}_i - \widehat{\tilde{\mu}}\right)\left(\tilde{X}_i^T - \widehat{\tilde{\mu}}\right)^T - I_p\right)\\
&= \sqrt{n}\ \mathrm{vec}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i^T - I_p\right) + \sqrt{n}\mathrm{vec}(\widehat{\tilde{\mu}}\ \widehat{\tilde{\mu}}^T);
\end{aligned}
$$

The normalized sum $\sqrt{n}\ \mathrm{vec}\left(\frac{1}{n}\sum_i \tilde{X}_i\tilde{X}_i^T - I_p\right)$ is asymptotically normal by the central limit theorem: the limiting covariance exists because we assume P_0 has finite fourth moments. Of course, $\widehat{\tilde{\mu}} = O_P(n^{-1/2})$ since it is the sample mean of the whitened data; this implies $\sqrt{n}\ \widehat{\tilde{\mu}}\ \widehat{\tilde{\mu}}^T = o_P(n^{-1/2})$. Thus $\sqrt{n}\mathrm{vec}\left(\widehat{\tilde{\Sigma}} - I_p\right)$ is asymptotically normal. The derivative of the transformation $\Sigma \to \Sigma^{-1/2}$ at the identity is $-\frac{1}{2}I_p$. By the Delta Method ([40], Theorem 3.1),

$$
\mathrm{vec}(\widehat{\tilde{\Sigma}}^{-1/2} - I_p) = -\frac{1}{2}\mathrm{vec}(\widehat{\tilde{\Sigma}} - I_p) + o_P(n^{-1/2}).
$$

Conclude:

$$\boxed{\begin{aligned}
&\left[\frac{1}{n^2}\sum_{i,j}\nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_i,\tilde{X}_j,\tilde{\Gamma}_0)\right]\left(\widehat{\Sigma}^{-1/2}-\Sigma_0^{-1/2}\right)\Sigma_0^{1/2}\tilde{X}_i\\
+\;&\left[\frac{1}{n^2}\sum_{i,j}\nabla_{\mathrm{vec}(B)}\nabla_y^T r(\tilde{X}_i,\tilde{X}_j,\tilde{\Gamma}_0)\right]\left(\widehat{\Sigma}^{-1/2}-\Sigma_0^{-1/2}\right)\Sigma_0^{1/2}\tilde{X}_j\\
=\;&-\frac{1}{2}\mathbb{E}_{\tilde{P}_0}\left[(\tilde{X})^T\otimes\nabla_x^T\mathrm{vec}\left(\psi(\tilde{X},\tilde{\Gamma}_0,\tilde{P}_0)\right)\right]\left(\Sigma_0^{\frac{1}{4}}\otimes\Sigma_0^{-\frac{1}{4}}\right)\\
&\times\mathrm{vec}\left(\frac{1}{n}\sum_{i=1}^n\tilde{X}_i\tilde{X}_i^T-I_p\right)+o_P(n^{-1/2}).
\end{aligned}}$$

Show:

$$\begin{aligned}
&\left[\frac{1}{n^2}\sum_{i,j}\nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_i,\tilde{X}_j,\tilde{\Gamma}_0)\right]\Sigma_0^{-1/2}(\mu_0-\hat{\mu})\\
+\;&\left[\frac{1}{n^2}\sum_{i,j}\nabla_{\mathrm{vec}(B)}\nabla_y^T r(\tilde{X}_j,\tilde{X}_j,\tilde{\Gamma}_0)\right]\Sigma_0^{-1/2}(\mu_0-\hat{\mu})\\
\approx&-\mathbb{E}_{\tilde{P}_0}\left[\nabla_x^T\mathrm{vec}\left(\psi(\tilde{X},\tilde{\Gamma}_0,\tilde{P}_0)\right)\right]\left(\frac{1}{n}\sum_{i=1}^n\tilde{X}_i\right)
\end{aligned}$$

The quantity $\Sigma_0^{-1/2}(\hat{\mu}-\mu_0)$ is precisely $\frac{1}{n}\sum_{i=1}^n\tilde{X}_i$, the sample mean of the whitened data. The term is also order $O_P(n^{-1/2})$. So if we can replace the random matrices with their expectations, we pay an error of of $o_P(n^{-1/2})$. What we will show is:

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i,j}\nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_i,\tilde{X}_j,\tilde{\Gamma}_0)+\frac{1}{n^2}\sum_{i,j}\nabla_{\mathrm{vec}(B)}\nabla_y^T r(\tilde{X}_i,\tilde{X}_j,\tilde{\Gamma}_0)\\
&\qquad\qquad\qquad\qquad\xrightarrow{P}\mathbb{E}_{\tilde{P}_0}\left[\nabla_x^T\mathrm{vec}\left(\psi(\tilde{X},\tilde{\Gamma}_0,\tilde{P}_0)\right)\right].
\end{aligned}$$

The convergence of $\frac{1}{n^2}\sum_{i,j}\nabla_{\mathrm{vec}(B)}\nabla_x^T r(\tilde{X}_i,\tilde{X}_j,\tilde{\Gamma}_0)$ and $\frac{1}{n^2}\sum_{i,j}\nabla_{\mathrm{vec}(B)}\nabla_y^T r(\tilde{X}_i,\tilde{X}_j,\tilde{\Gamma}_0)$ to their respective expectations is established by the U-statistic theory we have been repeatedly using. And we have already derived the formula for the gradient of ψ with respect to its first argument:

$$\nabla_x^T \mathrm{vec}(\psi(x, \tilde{\Gamma}_0, \tilde{P}_0)) = \mathbb{E}_{\tilde{P}_0}\left[\nabla_{\mathrm{vec}(B)} \nabla_x^T r(x, X, \tilde{\Gamma}(B))\right] + \mathbb{E}_{\tilde{P}_0}\left[\nabla_{\mathrm{vec}(B)} \nabla_y^T r(X, x, \Gamma(\tilde{B}))\right].$$

Convergence, in conjunction with this formula, establish:

$$\boxed{\begin{aligned}
&\left[\frac{1}{n^2}\sum_{i,j} \nabla_{\mathrm{vec}(B)} \nabla_x^T r(\tilde{X}_i, \tilde{X}_j, \tilde{\Gamma}_0)\right] \Sigma_0^{-1/2}(\mu_0 - \hat{\mu})\\
+ &\left[\frac{1}{n^2}\sum_{i,j} \nabla_{\mathrm{vec}(B)} \nabla_y^T r(\tilde{X}_j, \tilde{X}_j, \tilde{\Gamma}_0)\right] \Sigma_0^{-1/2}(\mu_0 - \hat{\mu})\\
= &- \mathbb{E}_{\tilde{P}_0}\left[\nabla_x^T \mathrm{vec}\left(\psi(\tilde{X}, \tilde{\Gamma}_0, \tilde{P}_0)\right)\right]\left(\frac{1}{n}\sum_{i=1}^n \tilde{X}_i\right) + o_P(n^{-1/2}).
\end{aligned}}$$

Expansion of $\frac{1}{n^2}\sum_{i,j} \nabla_{\mathrm{vec}(B)} r(\tilde{X}_i, \tilde{X}_j, \widehat{\tilde{\Gamma}}_n)$.

Since $\widehat{\tilde{B}}_n = o_P(1)$ and the $\tilde{X}_i$ have mean 0 and identity covariance, the term $\frac{1}{n^2}\sum_{i,j} \nabla_{\mathrm{vec}(B)} r(\tilde{X}_i, \tilde{X}_j, \widehat{\tilde{\Gamma}}_n)$ has the same asymptotic expansion as in Theorem 2.3.4 (recall: $\tilde{F}$ is the distribution of $\tilde{\Gamma}_0^T \tilde{X}_1$)

$$\boxed{\begin{aligned}
&\frac{1}{n^2}\sum_{i,j} \nabla_{\mathrm{vec}(B)} r(\tilde{X}_i, \tilde{X}_j, \widehat{\tilde{\Gamma}}_n) = \frac{1}{n}\sum_{i=1}^n \mathrm{vec}(\psi(\tilde{X}_i, \tilde{\Gamma}_0, \tilde{P}_0))\\
+ &\frac{2}{3}\left(\frac{1}{3}\right)^{\frac{p-d}{2}} \left[M(\tilde{F}) \otimes I_{p-d}\right] \mathrm{vec}\widehat{\tilde{B}}_n + o_P(\|\widehat{\tilde{B}}_n\|_F) + o_P(n^{-1/2}),
\end{aligned}}$$

Asymptotic approximation of $\widehat{\tilde{B}}_n$.

We have the following asymptotic expansion:

$$
\begin{aligned}
0 = {} & \frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}(\psi(\tilde{X}_i,\tilde{\Gamma}_0,\tilde{P}_0)) + \frac{2}{3}\left(\frac{1}{3}\right)^{\frac{p-d}{2}}\left[M(\tilde{F})\otimes I_{p-d}\right]\mathrm{vec}(\widehat{\tilde{B}}_n) \\
- {} & \frac{1}{2}\mathbb{E}_{\tilde{P}_0}\left[(\tilde{X})^T\otimes\nabla_x^T\mathrm{vec}\left(\psi(\tilde{X},\tilde{\Gamma}_0,\tilde{P}_0)\right)\right]\left(\Sigma_0^{\frac{1}{4}}\otimes\Sigma_0^{-\frac{1}{4}}\right)\mathrm{vec}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i^T - I_p\right) \\
- {} & \mathbb{E}_{\tilde{P}_0}\left[\nabla_x^T\mathrm{vec}\left(\psi(\tilde{X},\tilde{\Gamma}_0,\tilde{P}_0)\right)\right]\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\right) \\
+ {} & o_P(n^{-1/2}) + o_P(\|\widehat{\tilde{B}}_n\|_F) \\
= {} & \frac{2}{3}\left(\frac{1}{3}\right)^{\frac{p-d}{2}}\left[M(\tilde{F})\otimes I_{p-d}\right]\mathrm{vec}(\widehat{\tilde{B}}_n) \\
+ {} & \frac{1}{n}\sum_{i=1}^{n}\Bigg\{\mathrm{vec}(\psi(\tilde{X}_i,\tilde{\Gamma}_0,\tilde{P}_0)) - \mathbb{E}_{\tilde{P}_0}\left[\nabla_x^T\mathrm{vec}\left(\psi(\tilde{X},\tilde{\Gamma}_0,\tilde{P}_0)\right)\right]\tilde{X}_i \\
& -\frac{1}{2}\mathbb{E}_{\tilde{P}_0}\left[(\tilde{X})^T\otimes\nabla_x^T\mathrm{vec}\left(\psi(\tilde{X},\tilde{\Gamma}_0,\tilde{P}_0)\right)\right]\left(\Sigma_0^{\frac{1}{4}}\otimes\Sigma_0^{-\frac{1}{4}}\right)\mathrm{vec}(\tilde{X}_i\tilde{X}_i^T - I_p)\Bigg\} \\
+ {} & o_P(n^{-1/2}) + o_P(\|\widehat{\tilde{B}}_n\|_F) \\
= {} & \frac{2}{3}\left(\frac{1}{3}\right)^{\frac{p-d}{2}}\left[M(\tilde{F})\otimes I_{p-d}\right]\mathrm{vec}(\widehat{\tilde{B}}_n) \\
+ {} & \frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}\left(\tilde{\psi}(\tilde{X}_i,\tilde{\Gamma}_0,\tilde{P}_0|\mu_0,\Sigma_0)\right) + o_P(n^{-1/2}) + o_P(\|\widehat{\tilde{B}}_n\|_F),
\end{aligned}
$$

where $\tilde{\psi}\,(x,\Gamma,P|\mu,\Sigma)$ was defined in the statement of Theorem 2.3.8. We can multiply the above expression through by $\left[M(\tilde{F})\otimes I_{p-d}\right]^{-1} = M(\tilde{F})^{-1}\otimes I_{p-d}$ without changing any stochastic order symbols: the matrix is deterministic, invertible, and its dimensions are fixed. Therefore we have:

$$
\begin{aligned}
\mathrm{vec}(\widehat{\tilde{B}}_n) + o_P(\|\widehat{\tilde{B}}_n\|_F) = {} & \frac{3}{2}3^{\frac{p-d}{2}}\left[M(\tilde{F})^{-1}\otimes I_{p-d}\right]^{-1} \\
& \times\frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}\left(\tilde{\psi}\left(\tilde{X}_i,\tilde{\Gamma}_0,\tilde{P}_0|\mu_0,\Sigma_0\right)\right) + o_P(n^{-1/2}).
\end{aligned}
$$

To show $\widehat{\tilde{B}}_n = O_P(n^{-1/2})$ take the norm of both sides:

$$\|\widehat{\tilde{B}}_n\|_F\big|1+o_P(1)\big| = \left\|\frac{3}{2}3^{\frac{p-d}{2}}\left[M(\tilde{F})^{-1}\otimes I_{p-d}\right]\frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}\left(\tilde{\psi}\left(\tilde{X}_i,\tilde{\Gamma}_0,\tilde{P}_0\big|\mu_0,\Sigma_0\right)\right)\right\|_2\big|1+o_P(1)\big|.$$

Clearly $\frac{3}{2}3^{\frac{p-d}{2}}\left[M(\tilde{F})^{-1}\otimes I_{p-d}\right]\frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}\left(\tilde{\psi}\left(\tilde{X}_i,\tilde{\Gamma}_0,\tilde{P}_0\big|\mu_0,\Sigma_0\right)\right) = O_P(n^{-1/2})$ by the multivariate central limit theorem. This establishes $\widehat{\tilde{B}}_n = O_P(n^{-1/2})$, which means the asymptotic expansion of $\widehat{\tilde{B}}_n$ can be written as a vector as in the following:

$$\mathrm{vec}(\widehat{\tilde{B}}_n) = \frac{3}{2}3^{\frac{p-d}{2}}\left[M(\tilde{F})^{-1}\otimes I_{p-d}\right]\frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}\left(\tilde{\psi}\left(\tilde{X}_i,\tilde{\Gamma}_0,\tilde{P}_0\big|\mu_0,\Sigma_0\right)\right) + o_P(n^{-1/2});$$

or as a $(p-d)\times d$ matrix, as in:

$$\widehat{\tilde{B}}_n = \frac{3}{2}3^{\frac{p-d}{2}}\frac{1}{n}\sum_{i=1}^{n}\tilde{\psi}\left(\tilde{X}_i,\tilde{\Gamma}_0,\tilde{P}_0\big|\mu_0,\Sigma_0\right)M(\tilde{F})^{-1} + o_P(n^{-1/2}).$$

This completes the proof.

□

B.3 Miscellaneous Proofs

B.3.1 A basic inequality.

Some proofs rely on a simple inequality satisfied by the criterion function ρ. Recall:

$$\rho(\Gamma,P) = \iint\left|\mathcal{X}(\Gamma s+\Gamma_\perp t;P) - \exp(-\|t\|_2^2/2)\mathcal{X}(\Gamma s;P)\right|^2\phi_d(s)\phi_{p-d}(t)\mathrm{d}s\mathrm{d}t,$$

where $\mathcal{X}(t;P) = \mathbb{E}_P\left(e^{it^TX}\right)$ is the characteristic function for the distribution P at t, Γ is a $p\times d$ orthogonal matrix, $\Gamma_\perp$ is any $p\times(p-d)$ orthogonal matrix that satisfies $\Gamma_\perp^T\Gamma = 0$, and ϕ_d (resp. ϕ_{p-d}) is the d (resp. $p-d$) standard normal density.

Proposition B.3.1. *Let P_1 and P_2 be two probability distributions on $\mathbb{R}^p$, and let Γ_1 and Γ_2 be two orthogonal $p\times d$ matrices. Then:*

$$\begin{aligned}\big|\rho(\Gamma_1,P_1)-\rho(\Gamma_2,P_2)\big| \le 4\int\phi_d(s)\phi_{p-d}(t)\Big\{&\big|\mathcal{X}(\Gamma_1 s+\Gamma_{1_\perp}t;P_1)-\mathcal{X}(\Gamma_2 s+\Gamma_{2_\perp}t;P_2)\big|\\ &+\big|\mathcal{X}(\Gamma_1 s;P_1)-\mathcal{X}(\Gamma_2 s;P_2)\big|\Big\}\mathrm{d}s\mathrm{d}t.\end{aligned}$$

Proof. Note:

$$\rho(\Gamma_1, P_1) - \rho(\Gamma_2, P_2) = \iint \phi_d(s)\phi_{p-d}(t) \left\{ \left| \mathcal{X}(\Gamma_1 s + \Gamma_{1_\perp} t; P_1) - \exp(-\|t\|_2^2/2)\mathcal{X}(\Gamma_1 s; P_1) \right|^2 \right.$$
$$\left. - \left| \mathcal{X}(\Gamma_2 s + \Gamma_{2_\perp} t; P_2) - \exp(-\|t\|_2^2/2)\mathcal{X}(\Gamma_2 s; P_2) \right|^2 \right\} \mathrm{d}s\mathrm{d}t.$$

We prove a simple inequality for any four complex numbers a, b, c and d. Clearly $|a-b|^2 - |c-d|^2 = (|a-b| + |c-d|)(|a-b| - |c-d|)$. Therefore, by two applications of the triangle inequality, $\left||a-b|^2 - |c-d|^2\right| \leq (|a| + |b| + |c| + |d|)\,|a-c-(b-d)|$. Put $a = \mathcal{X}(\Gamma_1 s + \Gamma_{1_\perp} t; P_1)$, $b = \exp(-\|t\|_2^2/2)\mathcal{X}(\Gamma_1 s; P_1)$, $c = \mathcal{X}(\Gamma_2 s + \Gamma_{2_\perp} t; P_2)$ and $d = \exp(-\|t\|_2^2/2)\mathcal{X}(\Gamma_2 s; P_2)$ and use the fact the modulus of a characteristic function is bounded by 1 to obtain:

$$\begin{aligned}
&\left|\rho(\Gamma_1, P_1) - \rho(\Gamma_2, P_2)\right| \\
&\qquad \leq 4 \int \phi_d(s)\phi_{p-d}(t) \Big| \mathcal{X}(\Gamma_1 s + \Gamma_{1_\perp} t; P_1) - \mathcal{X}(\Gamma_2 s + \Gamma_{2_\perp} t; P_2) \\
&\qquad\qquad - \exp(-\|t\|_2^2/2)\left[\mathcal{X}(\Gamma_1 s; P_1) - \mathcal{X}(\Gamma_2 s; P_2)\right] \Big| \mathrm{d}s\mathrm{d}t \\
&\qquad \leq 4 \int \phi_d(s)\phi_{p-d}(t) \left\{ \left|\mathcal{X}(\Gamma_1 s + \Gamma_{1_\perp} t; P_1) - \mathcal{X}(\Gamma_2 s + \Gamma_{2_\perp} t; P_2)\right| \right. \\
&\qquad\qquad \left. + \left|\mathcal{X}(\Gamma_1 s; P_1) - \mathcal{X}(\Gamma_2 s; P_2)\right| \right\} \mathrm{d}s\mathrm{d}t.
\end{aligned}$$

□

B.3.2 Derivation of the alternate form of ρ.

Proof of Proposition 2.3.3. Recall the formula for $\rho(\Gamma, P)$, defined on $p \times d$ orthonormal matrices Γ and p-dimensional distributions P:

$$\rho(\Gamma, P) = \iint \left| \mathcal{X}(\Gamma s + \Gamma_\perp t; P) - \exp(-\|t\|_2^2/2)\mathcal{X}(\Gamma s; P) \right|^2 \phi_d(s)\phi_{p-d}(t)\mathrm{d}s\mathrm{d}t,$$

where $\Gamma_\perp$ is any $p \times (p-d)$ orthogonal matrix satisfying $\Gamma^T \Gamma_\perp = 0$, $\mathcal{X}(u; P) = \mathbb{E}_P\left[\exp(iu^T X)\right]$ is the characteristic function of P and ϕ_k is the standard normal density function in k dimensions, i.e. $\phi_k(z) = (2\pi)^{-k/2}\exp\left(-\|z\|^2/2\right)$. Expand the square:

$$\begin{aligned}&\iint \Bigg\{ \left|\mathcal{X}(\Gamma s + \Gamma_\perp t; P)\right|^2 - 2\exp\left(-\frac{1}{2}\|t\|_2^2\right)\mathcal{X}(\Gamma s + \Gamma_\perp t; P)\overline{\mathcal{X}(\Gamma s, P)} \\ &+ \exp(-\|t\|_2^2)\left|\mathcal{X}(\Gamma s; P)\right|\Bigg\}\phi_d(s)\phi_{p-d}(t)\mathrm{d}s\mathrm{d}t\end{aligned}$$

$$\begin{aligned}&= \iint \phi_d(s)\phi_{p-d}(t)\int \mathrm{d}P(x)\exp\left[i(\Gamma s + \Gamma_\perp t)^T x\right]\int \mathrm{d}P(y)\exp\left[-i(\Gamma s + \Gamma_\perp t)^T y\right]\mathrm{d}s\mathrm{d}t \\ &- 2\iint \phi_d(s)\phi_{p-d}(t)e^{\frac{-\|t\|_2^2}{2}}\int \mathrm{d}P(x)\exp\left[i(\Gamma s + \Gamma_\perp t)^T x\right]\int \mathrm{d}P(y)\exp\left[-i(\Gamma s)^T y\right]\mathrm{d}s\mathrm{d}t \\ &+ \iint \phi_d(s)\phi_{p-d}(t)\exp(-\|t\|_2^2)\int \mathrm{d}P(x)\exp\left[i(\Gamma s)^T x\right]\int \mathrm{d}P(y)\exp\left[-i(\Gamma s)^T y\right]\mathrm{d}s\mathrm{d}t\end{aligned}$$

Use Fubini's theorem to change the order of integration and combine like terms to obtain:

$$\begin{aligned}&\iint \mathrm{d}P(x)\mathrm{d}P(y)\int \phi_d(s)\exp\left[i\left(\Gamma^T(x-y)\right)^T s\right]\mathrm{d}s\int \phi_{p-d}(t)\exp\left[i\left(\Gamma_\perp^T(x-y)\right)^T t\right]\mathrm{d}t \\ &- 2\iint \mathrm{d}P(x)\mathrm{d}P(y)\int \phi_d(s)\exp\left[i\left(\Gamma^T(x-y)\right)^T s\right]\mathrm{d}s\int \phi_{p-d}(t)e^{\frac{-\|t\|_2^2}{2}}\exp\left[\left(\Gamma_\perp^T x\right)^T t\right]\mathrm{d}t \\ &+ \iint \mathrm{d}P(x)\mathrm{d}P(y)\int \phi_d(s)\exp\left[i\left(\Gamma^T(x-y)\right)^T x\right]\mathrm{d}s\int \phi_{p-d}(t)\exp(-\|t\|_2^2)\mathrm{d}t.\end{aligned}$$

We now evaluate the integrals over the Gaussian distribution. Recall that the characteristic function of the $\mathcal{N}_k(\mu, \Sigma)$ evaluated at $u \in \mathbb{R}^k$ is given by $\exp\left(i\mu^T u - u^T\Sigma u/2\right)$. Therefore:

$$\begin{aligned}\int \phi_d(s)\exp\left[i\left(\Gamma^T(x-y)\right)^T s\right]\mathrm{d}s &= \exp\left(-\frac{1}{2}(x-y)^T\Gamma\Gamma^T(x-y)\right) \\ &= \exp\left(-\frac{1}{2}\|\Gamma^T(x-y)\|_2^2\right).\end{aligned}$$

Similarly,

$$\int \phi_{p-d}(t)\exp\left[i\left(\Gamma_\perp^T(x-y)\right)^T t\right]\mathrm{d}t = \exp\left(-\frac{1}{2}\|\Gamma_\perp^T(x-y)\|_2^2\right).$$

Now we calculate:

$$\begin{aligned}\int \phi_{p-d}(t)\exp(-\|t\|_2^2)\mathrm{d}t &= (2\pi)^{-\frac{p-d}{2}}\int \exp\left(-\frac{3}{2}\|t\|_2^2\right)\mathrm{d}t \\ &= \left(\frac{1}{3}\right)^{\frac{p-d}{2}}\int \left(\frac{1}{3}\right)^{-\frac{p-d}{2}}\phi_{p-d}\left(\frac{t}{\sqrt{1/3}}\right)\mathrm{d}t.\end{aligned}$$

Observe that $(1/3)^{(p-d)/2}\phi_{p-d}(t/\sqrt{1/3})$ is the density of the $\mathcal{N}(0, I_{p-d}/3)$ distribution. Hence the value of the integral above is $(1/3)^{(p-d)/2}$. The remaining integral to compute is:

$$\begin{aligned}
&\int \phi_{p-d}(t) \exp\left(-\|t\|_2^2/2\right) \exp\left[\left(\Gamma_\perp^T x\right)^T t\right] \mathrm{d}t \\
&= \left(\frac{1}{2}\right)^{\frac{p-d}{2}} \int \left(\frac{1}{2}\right)^{-\frac{p-d}{2}} \phi_{p-d}\left(\frac{t}{\sqrt{1/2}}\right) \exp\left[\left(\Gamma_\perp^T x\right)^T t\right] \mathrm{d}t \\
&= \left(\frac{1}{2}\right)^{\frac{p-d}{2}} \exp\left(-\frac{1}{4}\|\Gamma_\perp^T x\|_2^2\right),
\end{aligned}$$

since $\exp(-\|u\|_2^2/4)$ is the characteristic function of the $\mathcal{N}(0, \frac{1}{2}I_{p-d})$ distribution at u. Putting everything together we obtain:

$$\begin{aligned}
&\iint \mathrm{d}P(x)\mathrm{d}P(y) \exp\left[-\frac{1}{2}\left(\|\Gamma^T(x-y)\|_2^2 + \|\Gamma_\perp^T(x-y)\|_2^2\right)\right] \\
&- 2\left(\frac{1}{2}\right)^{\frac{p-d}{2}} \iint \mathrm{d}P(x)\mathrm{d}P(y) \exp\left(-\frac{1}{2}\|\Gamma^T(x-y)\|_2^2\right) \exp\left(-\frac{1}{4}\|\Gamma_\perp^T x\|_2^2\right)
\end{aligned}$$

We simplify the first term using the identity $\|x\|_2^2 = \|\Gamma^T x + \Gamma_\perp^T x\|_2^2 = \|\Gamma^T x\|_2^2 + \|\Gamma_\perp^T x\|_2^2$, which holds because Γ and $\Gamma_\perp$ are orthonormal matrices which span orthogonal spaces (i.e. the matrix $(\Gamma \;\; \Gamma_\perp)$ is $p \times p$ and orthogonal); therefore, $\rho(\Gamma, P) = \iint \mathrm{d}P(x)\mathrm{d}P(y)r(x, y, \Gamma)$ where:

$$\begin{aligned}
r(x, y, \Gamma) = \exp&\left(-\frac{1}{2}\|x-y\|_2^2\right) \\
&+ \exp\left(-\frac{1}{2}\|\Gamma^T(x-y)\|_2^2\right)\left[-2\left(\frac{1}{2}\right)^{\frac{p-d}{2}} \exp\left(-\frac{1}{4}\|\Gamma_\perp^T x\|_2^2\right) + \left(\frac{1}{3}\right)^{\frac{p-d}{2}}\right].
\end{aligned}$$

□

B.3.3 Bounds on the derivatives of $r(x, y, \Gamma)$.

This next lemma was essential for proving Theorem 2.3.4.

Lemma B.3.2 (Bounds on derivatives of $r(x, y, \Gamma(B))$ in B.). *For $r(x, y, \Gamma)$ as defined in Proposition 2.3.3, consider the composition $r(x, y, \Gamma(B))$ where $\Gamma(B)$ is the mapping:*

$$\Gamma(B) = (\Gamma_0 \;\; \Gamma_{0_\perp}) \exp\left(\begin{bmatrix} 0 & -B \\ B & 0 \end{bmatrix}\right) J_{p,d}.$$

($J_{p,d}$ consists of the first d columns of the $p \times p$ identity matrix) Then for $l = 1, 2, 3$:

$$\left| \frac{\partial^l}{\partial B_{i_1} ... \partial B_{i_l}} r(x, y, \Gamma(B)) \right| \leq K_l(p, d) \left[\|x\|_2^l + \|y\|_2^l \right],$$

where $K_l(p, d)$ is a constant that depends only on the dimensions p and d and the number of derivatives l, and not on B or the choice of indices.

This result can be extended to higher order derivatives.

Proof. Recall the form of the function $r(x, y, \Gamma(B))$:

$$r(x, y, \Gamma(B)) = \exp\left(-\frac{1}{2}\|x - y\|_2^2\right)$$
$$+ \exp\left(-\frac{1}{2}\|\Gamma(B)^T(x - y)\|_2^2\right) \left[-2\left(\frac{1}{2}\right)^{\frac{p-d}{2}} \exp\left(-\frac{1}{4}\|\Gamma_\perp(B)^T x\|_2^2\right) + \left(\frac{1}{3}\right)^{\frac{p-d}{2}} \right].$$

Since we are taking derivatives in the parameter B we can ignore the term $\exp\left(-\frac{1}{2}\|x - y\|_2^2\right)$. We are going to prove the bound in the case $l = 3$. The application of our method for bounding other derivatives will be immediately evident. We hope that proving this one example gives the reader a sufficient insight into understanding the behavior of the partial derivatives of r.

From the form of the function $r(x, y, \Gamma(B))$, to bound its third derivatives, it's enough to bound the third derivatives of the functions $\exp\left(-\frac{1}{2}\|\Gamma(B)^T(x - y)\|_2^2\right)$ and $\exp\left(-\frac{1}{4}\|\Gamma_\perp(B)^T x\|_2^2\right)$. We do so now.

Derivatives of $\exp\left(-\frac{1}{2}\|\Gamma(B)^T(x - y)\|_2^2\right)$**.**

First, consider a real-valued generic function $g(B)$. Introduce the shorthand:

$$\dot{g}_i = \frac{\partial}{\partial B_i} g(B)$$
$$\ddot{g}_{ij} = \frac{\partial^2}{\partial B_i \partial B_j} g(B)$$
$$\dddot{g}_{ijk} = \frac{\partial^3}{\partial B_i \partial B_j \partial B_k} g(B).$$

(here we just assumed we can differentiate g however many times we require). We do not require the three indices i, j and k to be unique. The third partial derivatives of the function $\exp(g(B))$ have the form:

$$\frac{\partial^3}{\partial B_i \partial B_j \partial B_k} \exp(g(B)) = \exp(g(B)) \left(\dot{f}_i \dot{f}_j \dot{f}_k + \ddot{f}_{ij} \dot{f}_k + \ddot{f}_{ik} \dot{f}_j + \ddot{f}_{jk} \dot{f}_i + \dddot{f}_{ijk} \right).$$

Let $g(B) = -\frac{1}{2}\|\Gamma(B)^T(x-y)\|_2^2 = -\frac{1}{2}(x-y)^T\Gamma(B)\Gamma(B)^T(x-y)$. Then:

$$\begin{aligned}
\dot{g}_i &= -(x-y)^T\dot{\Gamma}_i(B)\Gamma(B)^T(x-y) \\
\ddot{g}_{ij} &= -(x-y)^T\ddot{\Gamma}_{ij}(B)\Gamma(B)^T(x-y) - (x-y)^T\dot{\Gamma}_i(B)\dot{\Gamma}_j^T(x-y) \\
\dddot{g}_{ijk} &= -(x-y)\dddot{\Gamma}_{ijk}(B)\Gamma(B)^T(x-y) - (x-y)\ddot{\Gamma}_{ij}(B)\dot{\Gamma}_k(B)^T(x-y) \\
&\quad - (x-y)\ddot{\Gamma}_{ik}(B)\dot{\Gamma}_j(B)^T(x-y) - (x-y)\dot{\Gamma}_i\ddot{\Gamma}_{jk}(B)^T(x-y).
\end{aligned}$$

where we applied the shorthand derivative notation to the map $\Gamma(B)$ element-wise. So the third derivatives of $\exp\left(-\frac{1}{2}\|\Gamma(B)^T(x-y)\|_2^2\right)$ in B consist of the exponent itself multipled by a sum, which consists of products of terms of the form given in the above display. After some careful checking and book-keeping, we see that any given term in this sum contains exactly three derivatives of $\Gamma(B)$. For instance, we see terms of the form:

$$(x-y)^T\ddot{\Gamma}_{ij}(B)\Gamma(B)^T(x-y)(x-y)^T\dot{\Gamma}_k(B)\Gamma(B)^T(x-y),$$

since there are a total of three derivatives of Γ being taken in that term. Because of the behavior of the derivatives of the map $\Gamma(B)$, it turns out that the largest terms are those having the most odd-numbered derivatives of Γ; they are products of three terms which consist of one derivative:

$$-(x-y)\dot{\Gamma}_i(B)\Gamma(B)^T(x-y)(x-y)^T\dot{\Gamma}_j(B)\Gamma(B)^T(x-y)(x-y)^T\dot{\Gamma}_k(B)\Gamma(B)^T(x-y).$$

To see why these are the largest terms, we now consider derivatives of $\Gamma(B)$. Recall $\Gamma(B)$ consists of the first d columns of the matrix:

$$\begin{pmatrix}\Gamma(B) & \Gamma_\perp(B)\end{pmatrix} = \begin{pmatrix}\Gamma_0 & \Gamma_{\perp 0}\end{pmatrix} \exp\left(\begin{bmatrix} 0 & -B^T \\ B & 0 \end{bmatrix}\right)$$

Compute the partial derivative of the matrix with respect to B_i:

$$\begin{aligned}
\begin{pmatrix}\frac{\partial}{\partial B_i}\Gamma(B) & \frac{\partial}{\partial B_i}\Gamma_\perp(B)\end{pmatrix} &= \begin{pmatrix}\Gamma_0 & \Gamma_{\perp 0}\end{pmatrix} \exp\left(\begin{bmatrix} 0 & -B^T \\ B & 0 \end{bmatrix}\right) \begin{bmatrix} 0 & -(\mathbf{1}_i^{(p-d)\times d})^T \\ \mathbf{1}_i^{(p-d)\times d} & 0 \end{bmatrix} \\
&= \left(\Gamma_\perp(B)\mathbf{1}_i^{(p-d)\times d} \quad -\Gamma(B)(\mathbf{1}_i^{(p-d)\times d})^T\right),
\end{aligned}$$

where $\mathbf{1}_i^{(p-d)\times d} = \frac{\partial}{\partial B_i} B$ is a matrix whose ith entry is equal to 1 and the rest are 0. Thus we have $\dot{\Gamma}_i(B) = \Gamma_\perp(B)\mathbf{1}_i^{(p-d)\times d}$; in other words, taking a derivative flips the direction of $\Gamma(B)$ to its orthogonal complement. Moreover, it's easy to check:

$$\ddot{\Gamma}_{ij}(B) = -\Gamma(B)(\mathbf{1}_i^{(p-d)\times d})^T \mathbf{1}_j^{(p-d)\times d},$$

and:

$$\dddot{\Gamma}_{ijk}(B) = -\Gamma_\perp(B)\mathbf{1}_i^{(p-d)\times d}(\mathbf{1}_j^{(p-d)\times d})^T \mathbf{1}_k^{(p-d)\times d};$$

in fact, odd derivatives flip the direction of $\Gamma(B)$ and even derivatives preserve it. The above formulas justify the bounds:

$$\begin{aligned}
\|\dot{\Gamma}_i(B)^T(x-y)\|_2 &\leq \|\Gamma_\perp(B)^T(x-y)\|_2 \\
\|\dddot{\Gamma}_{ijk}(B)^T(x-y)\|_2 &\leq \|\Gamma_\perp(B)^T(x-y)\|_2 \\
\|\ddot{\Gamma}_{ij}(B)^T(x-y)\|_2 &\leq \|\Gamma(B)^T(x-y)\|_2.
\end{aligned}$$

Therefore, we can upper bound the third derivatives of $\exp\left(-\frac{1}{2}\|\Gamma(B)^T(x-y)\|_2^2\right)$ in B by sums of terms having the form:

$$\exp\left(-\frac{1}{2}\|x-y\|_2^2\right)\|\Gamma(B)^T(x-y)\|_2^l \|\Gamma_\perp(B)^T(x-y)\|_2^m,$$

where l and m are integers determined by the fact we took 3 derivatives, but not determined by the indices i, j and k. Moreover, the number of terms in this sum is determined completely by the fact that we take three derivatives. This is key: we get an upper bound that holds for all choices of indices. Furthermore, we can upper bound this term by

$$K_{l,m}\left(\|x\|_2^m + \|y\|_2^m\right);$$

to see why, observe that the exponential function $f(x) = x^{k_1}\exp(-cx^{k_2})$ for positive c, k_1 and even integer k_2 is uniformly bounded on $\mathbb{R}$ by a constant that only depends on these constants. Then, we just use the crude upper bound $\|\Gamma_\perp(B)^T(x-y)\|_2 \leq \|x\|_2 + \|y\|_2$.

Remember that for a given term, m represents the number of odd-numbered derivatives in that term. Therefore $m \leq 3$ since, as we observed before, each term consists of exactly three derivatives, so there cannot be a greater number of odd-numbered derivatives than 3. As we mentioned, there is a term which does have 3 odd derivatives, the term which is the product of three single derivatives:

$$-(x-y)\dot{\Gamma}_i(B)\Gamma(B)^T(x-y)(x-y)^T\dot{\Gamma}_j(B)\Gamma(B)^T(x-y)(x-y)^T\dot{\Gamma}_k(B)\Gamma(B)^T(x-y).$$

The final form of the upper bound is:

$$\left| \frac{\partial^3}{\partial B_i \partial B_j \partial B_k} \exp\left(-\frac{1}{2}\|\Gamma(B)^T(x-y)\|_2^2\right) \right| \leq K_3 \left(\|x\|_2^3 + \|y\|_2^3\right).$$

Derivatives of $\exp\left(-\frac{1}{4}\|\Gamma_\perp(B)^T x\|_2^2\right)$**.**

Bounding the derivatives of this term entails exactly the same procedure. The odd derivatives of $\Gamma_\perp(B)$ are in the column space of Γ, while the even derivatives remain in the column space of $\Gamma_\perp$. The exponential function $\exp\left(-\frac{1}{4}\|\Gamma_\perp(B)^T x\|_2^2\right)$ can uniformly bound the even derivatives but not the odd derivatives. Therefore:

$$\left| \frac{\partial^3}{\partial B_i \partial B_j \partial B_k} \exp\left(-\frac{1}{4}\|\Gamma_\perp(B)^T x\|_2^2\right) \right| \leq K_3' \|x\|_2^3.$$

Derivatives of $\exp\left(-\frac{1}{2}\|\Gamma(B)^T(x-y)\|_2^2 - \frac{1}{4}\|\Gamma_\perp(B)^T x\|_2^2\right)$**.**

It is not very hard to use our bounding method to show:

$$\left| \frac{\partial^l}{\partial B_{i_1} \dots \partial B_{i_l}} \exp\left(-\frac{1}{2}\|\Gamma(B)^T(x-y)\|_2^2\right) \right| \leq K_l \left(\|x\|_2^l + \|y\|2^l\right),$$

for any positive integer l. The cases $l = 1$ and $l = 2$ are particularly simple. It is also simple to show:

$$\left| \exp\left(-\frac{1}{4}\|\Gamma_\perp(B)^T x\|_2^2\right) \right| \leq K_l' \|x\|_2^l$$

for integers l, including $l = 1, 2$. To bound the third derivatives of the product $\exp\left(-\frac{1}{2}\|\Gamma(B)^T(x-y)\|_2^2 - \frac{1}{4}\|\Gamma_\perp(B)^T x\|_2^2\right)$, observe that, for generic functions f and g for which the derivatives exist,

$$(fg)_{ijk} = \dddot{f}_{ijk} g + \ddot{f}_{ij}\dot{g}_k + \ddot{g}_{jk}\dot{f}_i + \dot{g}j\ddot{f}_{ik} + \ddot{f}_{jk}\dot{g}_j + \dot{f}_j\ddot{g}_{ik} + \dot{f}_k\ddot{g}_{ij} + f\dddot{g}_{ijk}.$$

Therefore, for some other constant K_3'' we have:

$$\left| \frac{\partial^3}{\partial B_i \partial B_j \partial B_k} \exp\left(-\frac{1}{2}\|\Gamma(B)^T(x-y)\|_2^2 - \frac{1}{4}\|\Gamma_\perp(B)^T x\|_2^2\right) \right| \leq K_3'' \left(\|x\|_2^3 + \|y\|_2^3\right).$$

The lemma is shown. □

This final Lemma is essential for proving Theorem 2.3.8. It is necessary for proving that certain remainders are negligible, and certain cross derivatives of the ψ function are integrable.

Lemma B.3.3 (Bounds on cross-derivatives of $r(x, y, \Gamma(B))$ in x, y and B.)**.** *Consider the function $r(x, y, \Gamma(B))$. Let $z = (x, y, B)$ be the concatenation of all arguments to the function. Then for some subcollection of three variables z_i, z_j and z_k we have:*

$$\left|\frac{\partial^3}{\partial z_i \partial z_j \partial z_k} r(x, y, \Gamma(B))\right| \leq K_m(\|x\|_2^m + \|y\|_2^m),$$

where m is the number of z_{i_j} such that $z_{i_j} = B_{i_k}$ for some index i_k, and K_m is a constant that depends on m and not on the choice of indices i, j and k.

The Lemma essentially says that we can bound any of the third derivatives of $r(x, y, \Gamma(B))$ by powers of $\|x\|_2$ and $\|y\|_2$ that are determined by how many partial derivatives are taken in B. We conjecture that this result holds beyond third derivatives, i.e.

$$\frac{\partial^m}{\partial x_{j_1} \ldots \partial x_{j_{m_1}} \partial y_{k_1} \ldots \partial y_{k_{m_2}} \partial B_{l_1} \ldots \partial B_{l_{m_3}}} r(x, y, \Gamma(B)) \leq K(p, d) \left(\|x\|_2^{m_3} + \|y\|_2^{m_3}\right).$$

However, of primary interest to us is to show that the results of Theorem 2.3.8 go through.

Proof. In Lemma B.3.2 we did the case of taking three partial derivatives in B. So in this lemma, we just consider the case of mixed third derivatives with no variables from B; with one variable from B; and with two variables from B.

Case: no derivatives in B variables.

This case is not particularly interesting from the point of view of Theorem 2.3.8, where each term involved a derivative in B. So we just make some general remarks. Any derivative of $\exp\left(-\frac{1}{2}\|\Gamma(B)^T(x-y)\|_2^2\right)$ in a x or y variable will return the exponent itself, times a linear form in $\Gamma(B)^T(x-y)$. The exponential always dominates these linear forms, and thus we can bound derivatives in x and y by constants. Since p is fixed, take the largest constant as a universal bound.

The same reasoning holds for the product function

$$\exp\left(-\frac{1}{2}\|\Gamma(B)^T(x-y)\|_2^2 - \frac{1}{4}\|\Gamma_\perp(B)^T x\|_2^2\right).$$

Case: one derivative in a B variable.

For the purposes of representing the derivatives, it is easiest to get all the x or all the y second derivatives simultaneously. So consider:

$$
\begin{aligned}
\nabla_x^2 r(x,y,\Gamma(B)) = \left(\frac{1}{3}\right)^{\frac{p-d}{2}} &\exp\left(-\frac{1}{2}\|\Gamma(B)^T(x-y)\|_2^2\right) \\
&\times\Big[-\Gamma(B)\Gamma(B)^T + \Gamma(B)\Gamma(B)^T(x-y)(x-y)^T\Gamma(B)\Gamma(B)^T\Big] \\
&-2\left(\frac{1}{2}\right)^{\frac{p-d}{2}} \exp\left(-\frac{1}{2}\|\Gamma(B)^T(x-y)\|_2^2 - \frac{1}{4}\|\Gamma_\perp(B)^T x\|_2^2\right) \\
\Big[-\Gamma(B)\Gamma(B)^T - \frac{1}{2}\Gamma_\perp(B)\Gamma(B_\perp)^T &+ \Gamma(B)\Gamma(B)^T(x-y)(x-y)^T\Gamma(B)\Gamma(B)^T \\
+\frac{1}{4}\Gamma_\perp(B)\Gamma_\perp(B)^T x x^T\Gamma_\perp\Gamma_\perp(B)(B)^T &+ \frac{1}{2}\Gamma_\perp(B)\Gamma_\perp(B)^T x(x-y)\Gamma(B)\Gamma(B)^T \\
&+\frac{1}{2}\Gamma(B)\Gamma(B)^T(x-y)x\Gamma_\perp(B)\Gamma_\perp(B)^T\Big].
\end{aligned}
$$

While this function looks daunting to bound, its structure is advantageous. It consists of exponential functions of $\|\Gamma(B)^T(x-y)\|_2$ and $\|\Gamma_\perp(B)^T x\|_2$ multiplied by linear functions of $\Gamma(B)^T(x-y)$ and $\Gamma_\perp(B)^T x$; these linear functions are uniformly bounded by the exponential terms. Now, taking a partial derivative in B_i, will, by the product rule, only "flip" one $\Gamma(B)$ to a $\dot{\Gamma}_i(B)$ (or a $\Gamma_\perp(B)$ to a $\dot{\Gamma}_{\perp i}(B)$) one at a time for each term. From the proof of Lemma B.3.2 we know that $\dot{\Gamma}_i(B)$ is in the direction of $\Gamma_\perp$ (and $\dot{\Gamma}_{\perp i}(B)$ is in the direction of Γ) so they may not be uniformly bounded by the exponential functions; instead, we use the simple bound $\|\Gamma_\perp(x-y)\|_2 \leq \|x\|_2 + \|y\|_2$. We do this at most once for each term in the above matrix, letting the exponential function bound the remaining terms. This produces an upper bound on all the derivatives of $\|x\|_2 + \|y\|_2$ (up to constants).

The matrices $\nabla_y^2 r(x,y,\Gamma(B))$ and $\nabla_x\nabla_y^T r(x,y,\Gamma(B))$ exhibit the same phenomenon. So third partial derivatives of r which include exactly one B variable are upper bounded by $K\left(\|x\|_2 + \|y\|_2\right)$.

Case: two derivatives in a B variable

We do the case where the third variable is an x variable by taking the gradient of $r(x,y,\Gamma(B))$ in x:

$$\begin{aligned}\nabla_x r(x,y,\Gamma(B)) = &\left(\frac{1}{3}\right)^{\frac{p-d}{2}} \exp\left(\frac{1}{2}\|\Gamma(B)^T(x-y)\|_2^2\right)\left[-\Gamma(B)\Gamma(B)^T(x-y)\right] \\ &- 2\left(\frac{1}{2}\right)^{\frac{p-d}{2}} \exp\left(-\frac{1}{2}\|\Gamma(B)^T(x-y)\|_2^2 - \frac{1}{4}\|\Gamma_\perp(B)^T x\|_2^2\right) \\ &\qquad \times \left[-\Gamma(B)\Gamma(B)^T(x-y) - \frac{1}{2}\Gamma_\perp(B)\Gamma_\perp(B)^T x\right].\end{aligned}$$

So we are dealing again with taking partial derivatives in B of exponential functions of $\|\Gamma(B)^T(x-y)\|_2$ and $\|\Gamma_\perp(B)^T x\|_2$ multiplied by linear functions of $\Gamma(B)^T(x-y)$ and $\Gamma_\perp(B)^T x$. Further, these linear functions are uniformly bounded by the exponent. Taking two partial derivatives with respect to B_i and B_j will only flip two instances of $\Gamma(B)$ (or $\Gamma_\perp(B)$) at a time. The exponential terms persist and bound the remaining terms uniformly, leaving a crude upper bound of $K\left(\|x\|_2^2 + \|y\|_2^2\right)$. The same techniques work for $\nabla_y r(x,y,\Gamma(B))$. This suffices to prove the lemma.

□