Multivariate Density Estimation
Comparing Transformation Random Forests, Triangular Transport Maps, and Copulas

---

Master's Thesis in Biostatistics (STA495)

by

Léon Kia Faro
13-795-026

supervised by

Prof. Dr. Torsten Hothorn

Zurich, September 2025

Abstract

This thesis evaluates three approaches to multivariate density estimation for tabular data within a single, consistent pipeline: separable triangular transport maps (TTM-Sep), Transformation Random Forests (TRTF; additive predictor), and copulas (used only for low-dimensional diagnostics, $K \leq 3$). All methods use standardized inputs and a common evaluation protocol so that likelihoods, diagnostics, and compute are directly comparable. In the configuration studied (additive predictor and monotone CDF smoothing), TRTF and TTM-Sep yield the same triangular-likelihood form, which enables like-for-like evaluation.

On Half-Moon ($n = 250$), mean joint negative log-likelihoods (NLL; lower is better) were 1.71 (TRTF), 1.93 (TTM-Sep), and 1.54 (copula). On a four-dimensional autoregressive generator they were 4.53, 5.66, and 5.45, respectively; permutation averages confirm order sensitivity for triangular maps. On MiniBooNE ($K = 43$; sum test log-likelihood), TRTF reached -30.01 under the standard preprocessing and training budget used here; published flow models report values around -12 to -16 under their settings. These numbers are not strictly comparable but indicate the relative accuracy of this configuration.

Overall, TRTF tends to lead within the separable family at low dimension, while higher-dimensional datasets expose the limits of separable structure. We report robustness checks (ordering), calibration diagnostics, and the numerical safeguards used, and we outline directions toward richer parameterizations within the same evaluation frame.

Contents

Chapter 1

Introduction

Multivariate density estimation supports likelihood-based modeling, anomaly detection, simulation, and decision making under uncertainty. Tabular datasets often combine moderate to high dimension with context-dependent conditional structure. Conditional variance can change with predictors, and conditional skewness or modality can depend on earlier variables. A transport perspective addresses these challenges by coupling the target to a simple reference through an invertible map. This perspective enables exact likelihoods, transparent conditionals, and efficient sampling through a shared computational backbone [Rosenblatt, 1952, Knothe, 1957]. This thesis compares three estimator families inside a single evaluation frame with matched objectives and diagnostics. The frame evaluates lower-triangular transport maps, Transformation Random Forests interpreted via probability integral transforms, and copulas that decouple marginals from dependence [Hothorn and Zeileis, 2017, Hothorn et al., 2018, Sklar, 1959].

We adopt a shared evaluation protocol summarized in Section 1.2 and detailed in Chapter 2; Figure A.1 in Appendix A shows the pipeline at a glance.

For clarity, efficiency, and interpretability we use separable triangular components as in Equation (2.6). This structure stabilizes the triangular determinant, enables exact inversion by back substitution, and fixes conditional shape across contexts.

1.1 Thesis and Problem Statement

This thesis investigates tabular multivariate density estimation within a unified transport-based evaluation frame. We compare separable triangular transport maps (TTM-Sep), Transformation Random Forests (TRTF), and copula baselines.

A transport perspective couples standardized data to a Gaussian reference through a monotone lower-triangular map. This structure yields exact likelihoods, transparent conditionals, exact inversion by back substitution, and linear per-sample evaluation. The Rosenblatt and Knothe rearrangements justify the triangular coupling for any variable order [Rosenblatt, 1952, Knothe, 1957].

We standardize features using training statistics only. Equation (2.1) defines the standardized coordinates u used for evaluation. All derivatives and Jacobians are computed in u. The diagonal affine correction in Equation (2.3) reports log densities on the original scale x. This convention keeps objectives, diagnostics, and comparisons interoperable across estimators and datasets. All log quantities are reported in nats.

We denote the K-variate standard normal density by η, and the univariate density and CDF by φ and Φ. Abbreviations for models and references appear in Table 1.1. Copulas decouple marginals from dependence and serve as interpretable baselines.

Separable triangular maps decompose each component into a context shift and a univariate monotone shape as in Equation (2.6). The decomposition fixes conditional shape across contexts and stabilizes the triangular determinant. Under strictly increasing conditional CDFs after standard monotone smoothing and with an additive predictor, TRTF implements the same separable triangular likelihood via the probability integral transform. Copulas preserve explicit marginals and model dependence on the unit hypercube; in this thesis they serve strictly as low-dimensional ($K \leq 3$) diagnostic baselines and are not evaluated on high-K datasets.

We evaluate all estimators under the single protocol referenced above, with matched preprocessing and reporting to keep results comparable. Section 3.3 defines metrics and timing conventions.

Figure A.1 in Appendix A visualizes the pipeline by showing standardization $u = T_{\text{std}}(x)$, the triangular transport branch containing TTM-Sep and TRTF, and the copula branch. Both branches feed the reported outputs, namely log density, conditionals, sampling, calibration, and compute, under the shared frame.

The central problem is to determine when separability is appropriate for tabular data. We study how TRTF and copulas position themselves against direct triangular transports inside the same reporting convention (reported log densities on x apply the affine correction in Eq. (2.3)). Ordering effects, conditional calibration, and computational trade-offs address this question.

On synthetic data, TRTF tends to outperform separable TTM variants yet shares their separability limits; on the MiniBooNE benchmark it improves on Gaussian references but trails published flow baselines. Chapter 3 presents the evidence and discusses these comparisons.

Non-goals. We do not treat high-capacity normalizing flows as primary models, and we restrict nonparametric copulas to low dimensions. We also exclude non-separable TRTF predictors and cross-term triangular maps because they raise compute and tuning costs substantially and complicate calibration/identifiability in our setting; the present scope focuses on separable structure for transparency and exact inversion. Section 1.3 states the scope and non-goals for reference. A brief pointer to expected changes if these variants were enabled—how cross-terms or non-additive predictors would likely affect calibration and NLL—is given in Chapter 4, Section 4.2.

1.2 The Transport Frame on One Page

To avoid duplication, the canonical derivations and notation live in Chapter 2. This section serves only as a map: we standardize with train-only statistics (Eq. (2.1)), evaluate likelihoods via the pullback (Eq. (2.2)), exploit the triangular determinant factorization (Eq. (2.5)), and apply the affine correction for reporting (Eq. (2.3)). Separable components are defined in Eq. (2.6). Figure A.1 in Appendix A illustrates the pipeline.

Notation remains consistent. We write η for the K-variate standard normal density, and φ and Φ for the univariate standard normal density and CDF. We reserve u for standardized coordinates and x for original coordinates, and we compute all derivatives with respect to u. These choices align symbols across Chapters 1–3 and prevent ambiguity in later diagnostics and tables.

This one-page frame removes duplicated exposition from Chapter 2. It establishes where logs and Jacobians live and makes complexity, inversion, and units explicit before the comparisons that follow. Section 1.1 documented the motivation, and Section 1.3 states the resulting contributions and research questions.

1.3 Contributions and Research Questions

This section states the contributions and the research questions, and maps them to the chapters and figures that deliver the evidence. We adopt the shared transport frame summarized in Section 1.2; Chapter 2 records notation and assumptions, and Figure A.1 anchors the comparisons.

The first contribution formalizes a unified likelihood view for separable triangular transport maps, Transformation Random Forests, and copula baselines. Where we claim an equivalence between TRTF and separable triangular transports, it holds under the conditions made explicit in Section 2.2.1 (strictly increasing conditional CDFs after monotone smoothing and an additive predictor). Chapter 2 and Figure A.1 establish the conventions and remove duplication in later chapters.

The second contribution provides empirical benchmarks under a single protocol with matched preprocessing and reporting. We evaluate TTM-Sep, TTM-Marg, and TRTF on synthetic generators and real tabular data; copulas are included only as low-dimensional ($K \leq 3$) diagnostic baselines (e.g., Half-Moon, 4D) and are not used in high-K studies. The protocol records three families of measurements: average test log-likelihoods, conditional diagnostics based on probability integral transforms, and compute indicators for training and per-sample evaluation. Section 3.3 defines the protocol, Section 3.4 presents the synthetic and autoregressive results, and Section 3.5 positions our measurements against published normalizing-flow baselines where appropriate.

The third contribution distills practical guidance from the unified frame and the benchmarks. We state operational choices that preserve comparability, highlight ordering sensitivity and separability limits, and summarize when copulas serve as informative baselines. Chapter 4 consolidates these points as actionable recommendations and records limitations that motivate richer param-

Table 1.1: Model abbreviations used throughout the thesis.

Label	Meaning
TTM-Marg	Marginal triangular transport map (per-dimension; no context)
TTM-Sep	Separable triangular transport map (additive: g_k shift + monotone h_k)
TRTF	Transformation Random Forests (axis-parallel splits)
True-Marg	Oracle marginal density
True-Joint	Oracle conditional joint density
Copula	Copula baseline (Gaussian or nonparametric)

eterizations or alternative predictors.

Two questions drive the empirical study and bind the contributions to specific measurements. The first question asks how TRTF compares with TTM-Sep and copula baselines on synthetic data. All estimators share the transport frame in this comparison. We answer by reporting average test negative log-likelihoods, conditional negative log-likelihood decompositions, and probability integral transform diagnostics, with timing summaries that quantify practical cost. Section 3.4 provides the corresponding tables and figures.

The second question asks how closely our TRTF results on real benchmarks approach the published performance of modern normalizing flows under the standard preprocessing. We answer by placing our test log-likelihoods beside reported numbers from the literature. The gaps are interpreted through the separable Jacobian constraint and compute profiles. Section 3.5 reports these comparisons, and Chapter 4 interprets their implications for model choice.

Scope and non-goals maintain focus and ensure reproducibility. We study separable triangular maps and TRTF with additive predictors and compute all derivatives and Jacobians in standardized coordinates. The map direction $S : u \to z$ remains fixed for evaluation and inversion. Copulas include Gaussian dependence and a low-dimensional nonparametric variant used strictly as a diagnostic baseline. We treat high-capacity flows as external references rather than primary models, and we do not evaluate non-additive TRTF variants or cross-term triangular maps in this chapter. Chapter 2 records the formal assumptions and notation. Section 3.1 details preprocessing, seeds, and reporting conventions that keep results comparable across datasets and estimators.

Taken together, these commitments make the comparisons interpretable, keep units and complexity explicit, and prepare the reader for the empirical evidence that answers the two questions under a single, transparent evaluation frame.

Chapter 2

Methodological Background

2.1 Transport Frame and Notation

This section fixes the standardized coordinate system, notation, and algebraic identities used throughout the thesis. The motivation and schematic in Figure A.1 housed in Appendix A remain valid; here we strip the exposition down to the formulas needed in later chapters. We summarize the standardized pullback likelihood, state the triangularity assumption, and record the Jacobian factorization that drives evaluation and inversion.

We work with observations on the original scale $x \in \mathbb{R}^K$. Training-split statistics define a fixed standardization map

$$u \;=\; T_{\mathrm{std}}(x) \;=\; (x-\mu) \oslash \sigma, \qquad \sigma_k > 0, \tag{2.1}$$

where μ and σ denote the empirical mean and standard deviation estimated on the training split and $\oslash$ denotes elementwise division. In words, we shift and rescale features once, using training data only, and keep all derivatives and Jacobians in u-space to avoid leakage and to ensure comparability across estimators.

The standardized density π_U is coupled to a simple reference through a monotone triangular map $S : u \mapsto z$. Throughout the thesis the reference is the K-variate standard normal density $\eta(z)$. The pullback identity then reads

$$\pi_U(u) \;=\; \eta(S(u)) \; \left|\det \nabla_u S(u)\right|, \tag{2.2}$$

which evaluates the reference at $S(u)$ and applies the exact volume correction given by the Jacobian determinant. Reporting log densities on the original scale requires only the diagonal affine correction implied by standardization,

$$\log \pi_X(x) \;=\; \log \pi_U(T_{\mathrm{std}}(x)) - \sum_{k=1}^{K} \log \sigma_k. \tag{2.3}$$

We therefore differentiate with respect to u, and we convert to x-scale only at reporting time.

The transport is assumed to be lower triangular and componentwise monotone,

$$S(u) \;=\; \big(S_1(u_1), S_2(u_{1:2}), \ldots, S_K(u_{1:K})\big), \qquad \partial_{u_k} S_k(u_{1:k}) > 0, \tag{2.4}$$

Table 2.1: Notation for the transport frame used in Chapters 2 and 3. All derivatives and Jacobians are taken with respect to u; log densities on x-space apply the affine correction in Equation (2.3).

Symbol	Meaning
$x \in \mathbb{R}^K$	Original features on the data scale
T_{std}	Standardization map using training (μ, σ)
$u = T_{\text{std}}(x)$	Standardized evaluation coordinates
$z \in \mathbb{R}^K$	Reference coordinates after transport
$S : u \mapsto z$	Monotone lower-triangular transport map
$\nabla_u S(u)$	Jacobian of S with respect to u
$\eta(z)$	K-variate standard normal density
$\varphi(t)$, $\Phi(t)$	Univariate standard normal density and CDF
π_U, π_X	Densities on u- and x-space, respectively
μ, σ	Training mean vector and positive scales
K	Dimension of the feature vector

so the Jacobian $\nabla_u S(u)$ is lower triangular. Its determinant factorizes into a sum of one-dimensional log derivatives,

$$\log \left| \det \nabla_u S(u) \right| = \sum_{k=1}^{K} \log \partial_{u_k} S_k(u_{1:k}). \tag{2.5}$$

The factorization yields $\mathcal{O}(K)$ evaluation cost per-sample, improves numerical stability, and guarantees global invertibility: strictly monotone diagonal derivatives let us recover x by solving K one-dimensional monotone equations in sequence, mirroring the Rosenblatt and Knothe rearrangements [Rosenblatt, 1952, Knothe, 1957].

Table 2.1 consolidates the notation used in this transport frame. All derivatives and Jacobians act on u; the affine correction (2.3) converts log densities back to x for reporting. The remainder of this chapter adopts this frame. Section 2.2 details the separable triangular parameterization used for direct transports. Section 2.2.1 shows how Transformation Random Forests induce the same triangular likelihood via the probability integral transform under the conditions stated there (strictly increasing conditional CDFs after monotone smoothing and an additive predictor). Section 2.3 places copulas in the same reporting convention.

2.2 Separable Triangular Maps and Transformation Random Forests as Transport

This section unifies separable triangular maps and Transformation Random Forests (TRTF) within the transport frame fixed in Section 2.1. Both estimators realize a monotone lower-triangular map $S : u \mapsto z$ that couples the standardized target to the Gaussian reference η. The use of triangular transports builds on modern measure-transport literature; see, for instance,

triangular transformations and their properties in Bogachev et al. [2005]. Figure A.1 in Appendix A illustrates the shared backbone and locates TTM-Sep and TRTF on the transport branch introduced in Chapter 1. We focus on shared likelihood identities, modeling assumptions, and limits of separability, and defer implementation details to Chapter 3 and Appendix A.

The goal is to state a single likelihood for both constructions, clarify what separability permits, and identify failure modes that motivate richer parameterizations. We do not pursue non-additive TRTF predictors, cross-term triangular maps, or ordering heuristics in this section; Chapter 3 evaluates those choices empirically and Appendix A documents routines and defaults.

We adopt the notation introduced in Section 2.1. Coordinates satisfy $u = T_{\text{std}}(x)$, the reference density is $\eta(z)$, and the pullback identity (2.2) gives $\pi_U(u) = \eta(S(u))\,|\det \nabla_u S(u)|$. The map is lower-triangular with strictly positive diagonal partial derivatives, which yields the sum decomposition in Equation (2.5). These conventions keep derivatives in u-space and apply the affine correction (2.3) only when reporting $\log \pi_X(x)$.

We restrict attention to separable triangular maps. Component k decomposes into a context shift and a univariate monotone shape,

$$S_k(u_{1:k}) \;=\; g_k(u_{1:k-1}) + h_k(u_k), \qquad \log \partial_{u_k} S_k(u_{1:k}) \;=\; \log h_k'(u_k), \tag{2.6}$$

which fixes context effects in g_k and reserves h_k for the one-dimensional marginal shape. Intuitively, earlier coordinates translate the location, while the conditional shape along u_k remains fixed across contexts. The Jacobian contribution depends only on u_k, which reduces per-sample evaluation cost and simplifies inversion.

Assumptions. Unless stated otherwise, we assume:

- *Lower-triangularity:* S has the structure in Eq. (2.4).
- *Strict monotone coordinates:* $\partial_{u_k} S_k(u_{1:k}) > 0$ for all k and all arguments.
- *Separable component:* Eq. (2.6) holds, so conditional shape along u_k is fixed across contexts.

Substituting the standard normal reference into (2.2) produces a separable objective,

$$\log \pi_U(u) \;=\; \sum_{k=1}^{K} \Big[\log \varphi\big(S_k(u_{1:k})\big) + \log h_k'(u_k) \Big], \tag{2.7}$$

where φ denotes the univariate standard normal density. Equation (2.7) splits the log density into a reference fit and an exact volume correction. In plain language, the model evaluates how Gaussian each transformed coordinate appears, then corrects for the local stretch induced by h_k. The same decomposition produces linear per-sample time in K and stable accumulation of log derivatives.

The negative log-likelihood per-sample takes the quadratic-plus-barrier form

$$\mathcal{L}(u) \;=\; \sum_{k=1}^{K} \Big[\tfrac{1}{2}\, S_k(u_{1:k})^2 - \log h_k'(u_k) \Big], \tag{2.8}$$

which follows because $\log \varphi(t) = -\frac{1}{2}t^2 - \frac{1}{2}\log(2\pi)$ and constants independent of the parameters drop out. Equation (2.8) pulls each component toward the reference while preventing degenerate derivatives through the log barrier. In practice we enforce $h'_k(u_k) > 0$ by construction and control tails with mild regularization; implementation choices appear in Chapter 3 and Appendix A.

Separable structure encodes clear modeling assumptions. Conditional variance, skewness, and modality do not change with the preceding coordinates once g_k shifts location. Consequently, separable maps can underfit heteroskedastic or multimodal conditionals, which manifests as U-shaped or inverted-U probability integral transform (PIT) diagnostics. Variable ordering also matters for finite bases because triangular transports are anisotropic, even though a Knothe–Rosenblatt rearrangement exists for any ordering [Rosenblatt, 1952, Knothe, 1957]. These caveats guide the robustness checks in Chapter 3.

2.2.1 Transformation Random Forests within the Transport Frame

Transformation Random Forests [Hothorn and Zeileis, 2017, Hothorn et al., 2018, Hothorn and Zeileis, 2021] fit into the same transport frame through the probability integral transform. Let $\widehat{F}_k(\cdot \mid u_{1:k-1})$ denote the strictly increasing conditional CDF returned by a TRTF for coordinate k. (In practice, forest CDFs can be stepwise; we assume a measurable, strictly increasing version after standard monotone smoothing so that inversion and derivatives are well-defined.) The induced triangular component is

$$S_k(u_{1:k}) \;=\; \Phi^{-1}\Big(\widehat{F}_k(u_k \mid u_{1:k-1})\Big), \tag{2.9}$$

which maps conditionals to standard normal margins. In plain language, TRTF predicts a conditional CDF, then the probit transform places the result on the Gaussian reference scale. Differentiating $\Phi\big(S_k(u_{1:k})\big) = \widehat{F}_k(u_k \mid u_{1:k-1})$ with respect to u_k yields

$$\widehat{\pi}_k(u_k \mid u_{1:k-1}) \;=\; \varphi\big(S_k(u_{1:k})\big)\, \partial_{u_k} S_k(u_{1:k}), \tag{2.10}$$

which is exactly the pullback factor in Equation (2.7). Summing over k recovers Equation (2.7) in standardized coordinates.

The additive-predictor TRTF used in this thesis yields a separable transport. Under the model

$$\widehat{F}_k(u_k \mid u_{1:k-1}) \;=\; \Phi\big(h_k(u_k) + g_k(u_{1:k-1})\big), \tag{2.11}$$

we obtain

$$S_k(u_{1:k}) \;=\; h_k(u_k) + g_k(u_{1:k-1}), \qquad \partial_{u_k} S_k(u_{1:k}) \;=\; h'_k(u_k), \tag{2.12}$$

so TRTF implements the same separable triangular likelihood as the direct parameterization in Equation (2.6). The map is monotone in u_k by construction, the Jacobian depends only on u_k, and inversion proceeds by back-substitution identical to the separable map. This equivalence underpins the empirical comparisons in Chapter 3.

The equivalence also clarifies limits. Additive TRTF predictors shift location but cannot alter conditional shape with context, which mirrors the separable constraint. Axis-aligned partitions

stabilize estimation, yet they do not remove residual multimodality when the conditional shape varies with $u_{1:k-1}$. These limits are visible in PIT diagnostics and conditional negative log-likelihood decompositions on synthetic studies.

We emphasize operational scope and supporting references. All derivatives and Jacobians are computed in standardized coordinates, evaluation uses the triangular pullback, and reported log densities on the original scale include the affine correction (2.3). Implementation details on basis choices for h_k, feature construction for g_k, regularization, derivative clipping, timing, and memory footprints appear in Chapter 3 and Appendix A, which also provides pseudo-code for both estimators. Figure A.1 in Appendix A visualizes how the TTM-Sep and TRTF branches share the same computational path from standardized data to reported likelihoods.

In summary, separable triangular maps and additive-predictor TRTF realize the same lower-triangular likelihood once the data are standardized and the conditional CDFs are strictly increasing after monotone smoothing. The shared structure yields exact likelihoods, exact inversion, transparent conditionals, and linear per-sample complexity, but it restricts context-dependent shape. Section 2.3 positions copulas within the same reporting convention to decouple marginals from dependence.

2.3 Copula Baselines

This section positions copulas within the unified transport frame and links their reported likelihoods to the evaluation conventions used for triangular maps and Transformation Random Forests. Copulas decouple marginal modeling from dependence modeling by pairing univariate marginals with a separate dependence density on the unit hypercube [Nelsen, 2006, Joe, 2014]. Figure A.1 in Appendix A displays the copula branch beside the triangular branch and highlights how both yield comparable reported log densities under the shared evaluation pipeline.

We begin with pseudo-observations built from training-split marginals. Let $\widehat{F}_k$ denote the strictly increasing empirical or smoothed CDF of X_k estimated on the training split. Define the pseudo-observations and their probit transform as

$$v_k \;=\; \widehat{F}_k(x_k), \qquad z_k \;=\; \Phi^{-1}(v_k), \tag{2.13}$$

which map each coordinate to $(0,1)$ and then to $\mathbb{R}$ through the probit function. In plain language, the marginals become uniform scores, and z records those scores on a Gaussian scale. Mid-ranks and clamping near $(0,1)$ stabilize the transformation in finite samples.

The copula representation combines marginal densities with a dependence factor. The joint log density on the original scale satisfies

$$\log \widehat{\pi}_X(x) \;=\; \sum_{k=1}^{K} \log \widehat{f}_k(x_k) + \log c(v_1, \ldots, v_K)\,, \tag{2.14}$$

where c denotes the copula density on $(0,1)^K$. Equation (2.14) separates the task into two parts: fit interpretable marginals and correct for dependence through $\log c$. The reported quantity

already lives on the original scale, so the affine correction in Equation (2.3) is unnecessary. Figure A.1 in Appendix A uses the equivalent shorthand $\log c(z)$ because dependence is evaluated through $z = \Phi^{-1}(v)$.

The independence baseline fixes a lower bound for dependence modeling. Setting $c \equiv 1$ yields

$$\log \widehat{\pi}_X^{\text{ind}}(x) \;=\; \sum_{k=1}^{K} \log \widehat{f}_k(x_k), \tag{2.15}$$

which treats coordinates as independent after marginal fitting. Chapter 3 uses this baseline as a reference point in evaluation tables and figures.

The Gaussian copula specifies elliptical dependence through a correlation matrix Σ. With $z = \Phi^{-1}(v)$, the copula density admits the closed form

$$c_\Sigma(v) \;=\; |\Sigma|^{-1/2} \exp\Big(-\tfrac{1}{2}\, z^\top (\Sigma^{-1} - I) z \Big), \tag{2.16}$$

which reduces dependence estimation to fitting Σ on the transformed scores. In plain language, the Gaussian copula bends the joint shape away from independence according to Σ while preserving the learned marginals.

A low-dimensional nonparametric variant avoids elliptical assumptions at small K. We fit a kernel density $\widehat{f}_Z$ on $z = \Phi^{-1}(v)$ and recover the copula density by

$$c(v) \;=\; \frac{\widehat{f}_Z\big(\Phi^{-1}(v)\big)}{\prod_{k=1}^{K} \varphi\big(\Phi^{-1}(v_k)\big)}, \tag{2.17}$$

which applies the change of variables from z back to v and yields a proper copula density. In words, the kernel density models the joint shape of the probit scores, and division by the product of standard normal densities restores the unit-cube scale. This approach is viable only for small K, where kernel density estimation remains accurate and stable. We implement this baseline via the `kdecopula` package [Nagler, 2017]. Chapter 3 employs it strictly as a diagnostic baseline.

The transport frame keeps reporting interoperable across modeling branches despite distinct parameterizations. Triangular maps and TRTF evaluate the pullback likelihood in standardized coordinates and apply the fixed affine correction (2.3) when mapping back to x. Copulas operate on x directly through Equation (2.14), yet Figure A.1 in Appendix A shows how the probit scores z maintain comparability with the Gaussian reference used above. This alignment keeps objectives and diagnostics consistent across Chapters 2 and 3.

Modeling choices and limits follow from the chosen copula. The Gaussian copula imposes elliptical dependence and may misrepresent tail behavior or localized asymmetry. The nonparametric variant mitigates these issues only at small dimension and sufficient sample size. The independence baseline provides a transparent reference when dependence is weak or data are scarce. These caveats motivate treating copulas as interpretable baselines rather than definitive high-dimensional models in the empirical study of Chapter 3. Sklar's theorem underlies all constructions above and formalizes the decoupling of marginals from dependence [Sklar, 1959].

Chapter 3

Data Analysis and Validation

This chapter turns the commitments of Chapters 1 and 2 into a practical modeling program. Our aim is to express three model families—triangular transport maps (TTM), transformation random forests (TRTF), and copulas—within a common transport framework so that likelihoods, calibration, and computational cost are directly comparable. Every method we study standardizes the data, learns a monotone triangular map to a simple reference, and evaluates Jacobians in the standardized space. That alignment keeps objectives, diagnostics, and reported log-densities interoperable.

Model abbreviations. We follow Table 1.1 for concise labels (TTM-Marg, TTM-Sep, TRTF, True-Marg/True-Joint, Copula) across tables and figures.

3.1 Datasets and Preprocessing

This section fixes data sources, generators, and preprocessing so likelihoods, calibration, and compute remain comparable across models. All estimators operate in standardized coordinates, evaluate Jacobians in that space, and report log densities on the original scale using the common affine correction. We keep a single triangular-map direction $S : u \to z$ across methods to avoid mixed objectives. To avoid symbol collisions, σ denotes standardization scales only; logistic gates are written logistic$(\cdot)$ throughout.

We standardize features with training-split statistics only. Equation (2.1) defines $u = T_{\text{std}}(x) = (x - \mu) \oslash \sigma$ with $\sigma_k > 0$. We evaluate $\log \pi_U(u)$ through the pullback identity in Equation (2.2), apply the triangular factorization from Equation (2.5), then convert to $\log \pi_X(x)$ using the diagonal correction in Equation (2.3). We report average test negative log-likelihoods (NLL) in nats. Negative per-dimension NLL values can occur because valid densities may exceed one on subdomains. Figure A.1 in Appendix A shows the standardized pipeline shared by transport maps, Transformation Random Forests, and copulas.

We use fixed train, validation, and test splits with proportions 0.60/0.20/0.20 unless a benchmark provides official splits. Synthetic studies report results for $n \in \{25, 50, 100, 250\}$ and use $n = 250$

Table 3.1: Configuration for the four-dimensional autoregressive generator used in the synthetic study. The beta and gamma coordinates are two-component mixtures with logistic gates; fixed parameters and gates match the prose above.

Coordinate	Distribution	Parameters / gate
X_1	Normal	$\mathcal{N}(0,1)$
X_2	Exponential	rate $= \lambda_0 = 1$
$X_3 \mid X_{1:2}$	Mixture Beta	Beta$(2.5, 5.0)$ / Beta$(5.0, 2.5)$; $w = \text{logistic}(\gamma_0 + \gamma_1 X_1 + \gamma_2(X_2 - 1))$, $\gamma =$
$X_4 \mid X_{1:3}$	Mixture Gamma	Gamma$(k_1{=}3, r_1{=}1 + 0.5X_2)$ / Gamma$(k_2{=}6, r_2{=}0.75 + 0.25X_2)$; $\tilde{w} =$ lo

for headline tables and figures; for real-data benchmarks we use N to denote the training budget. The canonical four-dimensional ordering is $(1, 2, 3, 4)$. Robustness to ordering is assessed by averaging over all $4! = 24$ permutations. We adopt the natural column order for real datasets. We fix seeds $\{11, 13, 17, 19, 23\}$ for data generation and model fitting, and we average repeated runs with standard errors to quantify stochastic variability. Figure 3.2 displays the 20% test split for the synthetic calibration study.

The Half-Moon dataset provides a curved, bimodal joint in $K = 2$. We draw a class $Y \sim$ Bernoulli(0.5), an angle $\Theta \sim \text{Unif}[0, \pi]$, and additive noise $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_2)$ with $\sigma = 0.10$. For $Y = 0$ we set $m(\Theta) = (\cos\Theta, \sin\Theta)$. For $Y = 1$ we set $m(\Theta) = (1 - \cos\Theta, -\sin\Theta + 0.5)$. The observed $X = m(\Theta) + \varepsilon$. The "True joint" oracle evaluates the mixture density by numerical quadrature over Θ with the known Gaussian noise, and the "True conditional" oracle conditions on Y. Figure 3.1 (p. 18) shows representative contour plots at $n = 250$, which align with this generator. Table 3.2 (p. 17) reports the corresponding NLLs.

The four-dimensional autoregressive generator combines Gaussian, exponential, beta, and gamma components to induce heteroskedasticity, skew, and conditional multimodality. The first coordinate is $X_1 \sim \mathcal{N}(0, 1)$. The second coordinate is independent $X_2 \sim \text{Exp}(\lambda_0)$ with rate $\lambda_0 = 1$. The third coordinate lies on $(0, 1)$ and is a context-gated mixture of two beta laws, $X_3 \mid X_{1:2} \sim w\,\text{Beta}(\alpha_1, \beta_1) + (1 - w)\,\text{Beta}(\alpha_2, \beta_2)$. We set $(\alpha_1, \beta_1) = (2.5, 5.0)$ and $(\alpha_2, \beta_2) = (5.0, 2.5)$. The mixing weight is $w = \text{logistic}(\gamma_0 + \gamma_1 X_1 + \gamma_2(X_2 - 1))$ with logistic$(\cdot)$ the logistic function and $(\gamma_0, \gamma_1, \gamma_2) = (0, 1.5, 1.0)$. The fourth coordinate is positive and conditionally heteroskedastic, $X_4 \mid X_{1:3} \sim \tilde{w}\,\text{Gamma}(k_1, r_1(X_2)) + (1 - \tilde{w})\,\text{Gamma}(k_2, r_2(X_2))$. We set shapes $(k_1, k_2) = (3, 6)$, rates $r_1(X_2) = 1 + 0.5X_2$ and $r_2(X_2) = 0.75 + 0.25X_2$, and gate $\tilde{w} = \text{logistic}(\delta_0 + \delta_1 X_1 + \delta_3(X_3 - 0.5))$ with $(\delta_0, \delta_1, \delta_3) = (0, 1.0, 3.0)$. The "True joint" baseline uses these closed-form conditionals to evaluate the exact joint density, while the "True marginal" baseline uses the corresponding univariate marginals and ignores dependence.

Mixture weights use the logistic gate $\text{logistic}(a) = \exp(a)/(1 + \exp(a))$, which coincides with the two-component softmax and therefore keeps probabilities in $(0, 1)$ that sum to one. For completeness, the general softmax takes a vector a and returns $\text{softmax}(a)_i = \exp(a_i)/\sum_j \exp(a_j)$. This normalization is essential for the beta and gamma mixtures because it translates linear predictors into valid probability weights while preserving differentiability.

Table 3.1 (p. 12) summarizes the mixture families, gates, and fixed parameters by dimension.

Tables 3.4 (p. 20) and 3.6 (p. 20) then summarize the permutation and sample-size studies used later in this chapter.

The MiniBooNE benchmark follows the published preprocessing to ensure comparability with flow-based baselines. We remove 11 outliers with value -1000, drop seven features with extreme mass at a single value, and retain $K = 43$ attributes. We use the fixed train, validation, and test splits from the benchmark, apply train-only standardization, and avoid any extra pruning of correlated features. We report all log-likelihoods in nats and retain the published naming for flow comparators in later tables. Section 3.5 records these steps and provides the dataset context. Table 3.7 reproduces the flow baselines that motivate our TRTF runs.

Additional UCI datasets appear only when we retain them for real-data context. POWER keeps household electricity attributes after jittering the minute-of-day encoding, dropping the calendar date and reactive-power column, and adding uniform noise to break ties. GAS keeps the `ethylene_CO` subset, treats the series as i.i.d., removes strongly correlated attributes, and retains an eight-dimensional representation. HEPMASS keeps only the positive class from the "1000" split and discards five variables with repeated values to avoid density spikes. These preprocessing steps follow the same train-only standardization and reporting conventions described above. Section 3.5 provides the corresponding background and positions these datasets within our evaluation.

All models use the same standardized frame and direction for evaluation, which keeps objectives, diagnostics, and reported quantities interoperable across triangular transports, TRTF, and copula baselines. This alignment is necessary for the conditional decompositions, probability integral transform (PIT) calibration checks, and compute summaries presented later in this chapter.

3.2 Models and Implementation

This section specifies the estimators and implementation details that keep likelihoods, calibration, and compute directly comparable across models. All estimators share the transport direction $S : u \to z$, operate in standardized coordinates, and report log densities on the original scale using the affine correction from Chapter 2. Figure A.1 in Appendix A and Table 2.1 summarize the shared pipeline and notation.

We implement separable lower-triangular transport maps denoted TTM-Sep. Component k decomposes into a context shift and a univariate monotone shape,

$$S_k(u_{1:k}) = g_k(u_{1:k-1}) + h_k(u_k), \qquad \partial_{u_k} S_k(u_{1:k}) = h_k'(u_k) > 0, \tag{3.1}$$

so the Jacobian contribution depends only on u_k. The structure yields linear per-sample complexity in K and exact inversion by back-substitution.

We minimize the Gaussian pullback objective induced by the shared reference,

$$\mathcal{L}(u) = \sum_{k=1}^{K} \Big[\tfrac{1}{2} S_k(u_{1:k})^2 - \log h_k'(u_k) \Big], \tag{3.2}$$

which follows from the change-of-variables identity in Equation (2.2) combined with the triangular determinant factorization in Equation (2.5). The quadratic term pulls the transformed coordinates toward the reference, and the log-derivative term prevents degenerate solutions. We solve the regularized problem with bound-constrained optimization and enforce monotone structure by construction.

We construct h_k with monotone one-dimensional bases that combine identity, integrated sigmoids, softplus-like edge terms, and integrated radial basis functions. Nonnegativity constraints on the derivative coefficients guarantee $h_k'(u_k) \geq 0$. We linearize tails to stabilize likelihoods as $|u_k|$ grows. Ridge penalties apply to all basis coefficients, and optional sparsity penalties shrink context shifts when multicollinearity inflates variance. During training and evaluation we clip log-derivatives to $[-H, H]$ to avoid numerical overflow in the Jacobian sum; the bound H is tuned on the validation split.

We build g_k from low-degree polynomial features of $u_{1:k-1}$ and drop predecessors whose inclusion does not improve validation likelihood. This pruning keeps $\nabla_u S(u)$ sparse and improves stability in small-sample regimes. Ordering matters for finite bases, so headline results use the natural variable order while robustness studies vary the order as described in Section 3.1. When heuristics are applied, we evaluate two candidates on the validation split—identity and a Cholesky-pivoted ordering with optional Gaussianization—and select the ordering with the lower validation NLL. Appendix A records how the ordering is stored and reapplied at prediction time.

We reference a cross-term variant, denoted TTM-X, only to delimit scope. The variant augments the separable component with low-rank interactions,

$$S_k(u_{1:k}) = g_k(u_{1:k-1}) + h_k(u_k) + \sum_{j<k} \alpha_{kj}\, q_j(u_j)\, r_k(u_k), \tag{3.3}$$

where q_j and r_k are monotone features and constraints ensure $\partial_{u_k} S_k(u_{1:k}) > 0$. We exclude TTM-X from headline tables because the interactions alter identifiability and complicate calibration. The definition clarifies the naming used in the synthetic analyses.

We implement Transformation Random Forests with additive predictors and denote the model TRTF. Implementations rely on the `partykit` toolkit for recursive partitioning [Hothorn et al., 2015] together with the transformation-model framework [Hothorn and Zeileis, 2017, Hothorn et al., 2018]. Let $\widehat{F}_k(\cdot \mid u_{1:k-1})$ denote the strictly increasing conditional CDF returned by the forest. The induced triangular component is

$$S_k(u_{1:k}) = \Phi^{-1}\big(\widehat{F}_k(u_k \mid u_{1:k-1})\big), \tag{3.4}$$

and differentiation yields $\varphi\big(S_k(u_{1:k})\big)\, \partial_{u_k} S_k(u_{1:k}) = \widehat{\pi}_k(u_k \mid u_{1:k-1})$. Under the additive predictor $\widehat{F}_k(u_k \mid u_{1:k-1}) = \Phi\big(h_k(u_k) + g_k(u_{1:k-1})\big)$ we obtain $S_k = h_k + g_k$ and $\partial_{u_k} S_k = h_k'(u_k)$, which matches Equation (3.1) exactly in the transport frame. Consequently TRTF shares the likelihood in Equation (3.2), inherits exact inversion, and differs operationally through forest training and aggregation.

We keep copulas as dependence baselines with explicit scope. We fit only low-dimensional nonparametric copulas for $K \leq 3$, using probit-transformed pseudo-observations and kernel density

estimation on the Gaussian scale before mapping back to the unit cube with the appropriate Jacobian. The independence baseline evaluates the product of fitted marginals. We omit a Gaussian copula from the main experiments to preserve consistency with the low-K nonparametric dependence analyzed in Section 3.5.

We adopt a single inversion and evaluation convention across estimators. Training, Jacobians, and conditional evaluations occur in standardized coordinates. Sampling draws $z \sim \mathcal{N}(0, I)$, applies S^{-1} by back-substitution, and converts to the original scale with the stored affine parameters. This convention prevents mixed objectives and keeps all reported quantities interoperable.

We ensure reproducibility and comparability with fixed seeds, cached standardization parameters, and shared reporting utilities. Appendix A provides pseudo-code for TRTF fitting and prediction, nonparametric copulas, marginal and separable triangular maps, and the shared transport core that implements ordering, bases, derivatives, and evaluation. The appendix also records optimizer choices, timing hooks, and object layouts used in the experiments.

3.3 Evaluation Metrics and Protocol

This section defines the metrics and procedures applied across all models so likelihoods, calibration, and compute remain directly comparable. We evaluate every estimator in standardized coordinates, apply the triangular determinant, and report log densities on the original scale using the affine correction from Chapter 2. Figure A.1 in Appendix A and Table 2.1 summarize the shared pipeline and notation.

We distinguish pointwise log density from dataset averages. The test log likelihood (LL) is the mean of $\log \hat{\pi}_X(x)$ over the test split, and the test negative log likelihood (NLL) is its negative. Tables note "LL (higher is better)" or "NLL (lower is better)" to avoid ambiguity. Reported log densities on the original scale equal the standardized quantity minus $\sum_k \log \sigma_k$ as given by Equation (2.3). Consequently, datasets with larger training scales introduce large constant offsets that the affine correction removes.

Triangular models exploit the separable pullback in standardized coordinates. With $u = T_{\mathrm{std}}(x)$, the log density decomposes as

$$\log \hat{\pi}_U(u) = \sum_{k=1}^{K} \Big[\log \varphi\big(S_k(u_{1:k})\big) + \log \partial_{u_k} S_k(u_{1:k}) \Big], \tag{3.5}$$

so the determinant factorization in Equation (2.5) yields linear per-sample cost in K. In plain language, the model checks how Gaussian each transformed coordinate looks, then adds the exact log-Jacobian contribution from its one-dimensional derivative. The affine correction in Equation (2.3) converts $\log \hat{\pi}_U$ to $\log \hat{\pi}_X$ for reporting.

We report per-dimension conditional NLLs for triangular models to localize error. For each coordinate,

$$\mathrm{NLL}_k = -\frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} \log \hat{\pi}\big(x_{ik} \mid x_{i,1:k-1}\big), \tag{3.6}$$

and the joint NLL equals $\sum_{k=1}^{K} \mathrm{NLL}_k$ by construction. Copulas lack a unique triangular factorization, so we report only their joint NLL. Negative per-dimension NLL values can occur because valid densities may exceed one on subdomains. These conventions align with Equations (2.10) and (2.12), which link separable transports and Transformation Random Forests inside the common frame under the additive-predictor and monotone-smoothing assumptions stated in Section 2.2.1.

Calibration assesses whether predictive probabilities align with empirical frequencies. For triangular models we form conditional probability integral transform (PIT) values $V_{ik} = \widehat{F}_k(u_{ik} \mid u_{i,1:k-1})$ on the test split and expect independent $\mathrm{Unif}(0,1)$ draws under correct calibration [Gneiting et al., 2007]. We summarize departures from uniformity with the Kolmogorov–Smirnov statistic $D_n = \sup_t |\widehat{F}_n(t) - t|$ and report the associated p-value [Massey, 1951]. We complement the scalar summary with brief PIT distribution descriptions when patterns recur across seeds. For copulas we assess marginal PITs and low-dimensional slices where dependence is transparent. Systematic U-shaped or inverted-U PIT indicates under- or over-dispersion and motivates richer parameterizations.

Compute metrics quantify practical cost alongside fit. We record wall-clock training time on the training split and per-sample evaluation time on the test split. Triangular transports scale linearly in K and approximately linearly in the number of basis functions. Transformation Random Forests scale with the number and depth of trees per conditional during training, while prediction remains linear after aggregation. Copula training is dominated by correlation estimation or kernel density fitting, followed by fast evaluation. We also track peak resident memory when caching affects runtime. All timings use the deterministic pipeline defined in Chapter 3 and are averaged over seeds with standard errors.

Protocol choices keep comparisons stable and reproducible:

1. Standardize with training-split statistics, fit a single map $S : u \to z$, and evaluate Jacobians in standardized space.

2. Compute LL, NLL, and conditional decompositions in standardized coordinates, then apply the affine correction once for reporting.

3. Evaluate PIT diagnostics, Kolmogorov–Smirnov statistics, and compute metrics on the fixed test split with seeds $\{11, 13, 17, 19, 23\}$ and quote means with $\pm$ two standard errors across seeds (SE $= s/\sqrt{m}$ over m seeds).

Appendix A lists routine interfaces that support exact re-execution; figure captions and table notes repeat units, splits, and the "higher/lower is better" convention for clarity.

Numerical safeguards prevent unstable likelihoods from dominating summaries. We enforce strictly monotone derivatives by construction and clip log-derivative contributions to $[-H, H]$ during training and evaluation. We tune H and regularization on the validation split and reuse the selected bound on the test split. We report clipping status inline per study and keep the exact bound values with the corresponding experiment logs to avoid duplication. This practice

controls overflow in the Jacobian sum without masking systematic misfit that PIT diagnostics would reveal.

Scope limits clarify non-goals. We do not report AIC or BIC because effective parameter counts are not comparable across estimators in this frame. We also do not adjust p-values for multiple PIT checks; instead, we treat Kolmogorov–Smirnov results as diagnostics and corroborate them with effect sizes and plots. These limits keep the evaluation focused on likelihood, calibration, and compute under a single, transparent protocol.

3.4 Synthetic Results and Diagnostics

This section reports synthetic results for the Half-Moon and four-dimensional generators under the protocol in Section 3.3. We summarize mean test negative log likelihoods, per-dimension conditional NLLs, calibration evidence, and ordering robustness, referencing the corresponding tables and figures.

The Half-Moon generator stresses conditional shape in two dimensions. Table 3.2 lists mean joint NLLs with ± two standard errors: TRTF achieved 1.71 ± 0.09 nats, TTM-Sep achieved 1.93 ± 0.08 nats, and TTM-Marg achieved 2.02 ± 0.07 nats. The copula baseline reached 1.54 ± 0.09 nats and bracketed the triangular transports. The oracle references set 0.78 ± 0.10 nats for the true marginal density and 0.70 ± 0.12 nats for the true joint. Per-dimension NLLs confirm that the first coordinate is harder: TRTF reported $(1.23, 0.47)$, while TTM-Sep reported $(1.28, 0.65)$. Figure 3.1 shows contours consistent with these rankings and with the standardized pipeline in Figure A.1 of Appendix A. Clipping status: not triggered in these runs (no log-derivative terms reached the bound).

(mean NLL in nats).

Table 3.2: Half-Moon ($n = 250$): mean test negative log-likelihood (NLL; nats; lower is better). Values are means ± 2SE.

Model	Mean joint NLL	Conditional NLL 1	Conditional NLL 2
True-Marg	0.78 ± 0.10	0.39	0.39
True-Joint	0.70 ± 0.12	0.35	0.35
TRTF	1.71 ± 0.09	1.23	0.47
TTM-Marg	2.02 ± 0.07	1.28	0.74
TTM-Sep	1.93 ± 0.08	1.28	0.65
Copula	1.54 ± 0.09	0.77	0.77

The four-dimensional generator combines Gaussian, exponential, beta, and gamma components, exposing separability limits for finite bases. Table 3.3 (p. 18) reports the canonical ordering $(1, 2, 3, 4)$. TRTF aligned closely with the exponential coordinate, recording 1.51 nats compared with 1.49 for the true joint reference. TTM-Sep over-penalized that coordinate at 1.88 nats, and TTM-Marg overfit at 2.57 nats. The beta coordinate yielded negative NLLs for the oracles because valid densities can exceed one on $(0, 1)$; values were -0.79 for the true joint and -0.48

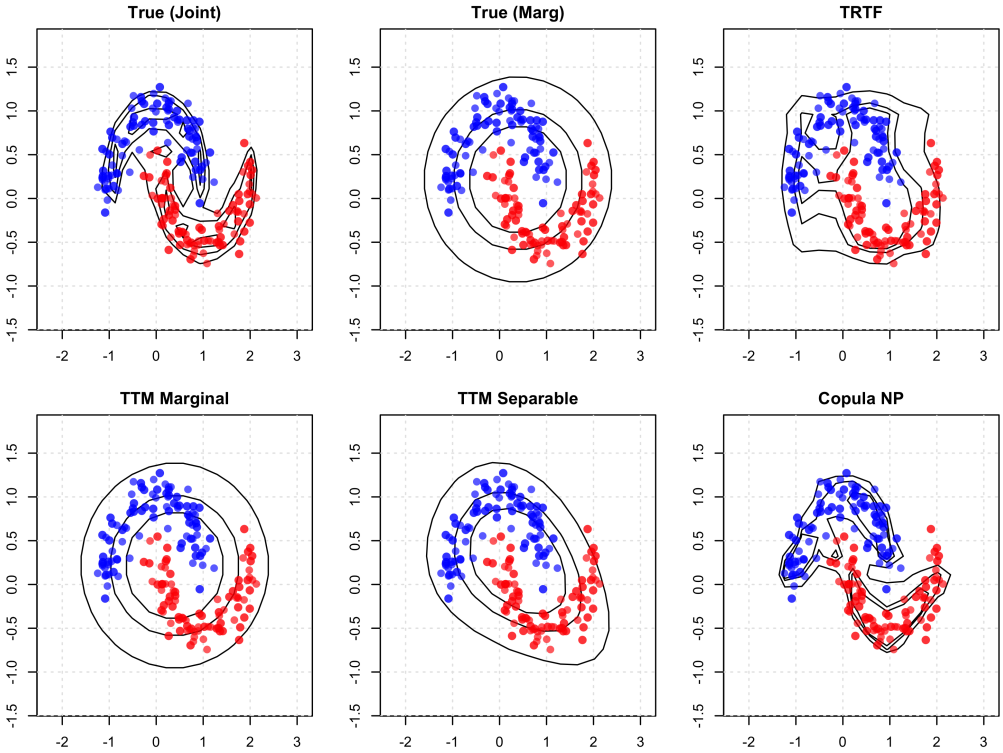

Figure 3.1: Half-Moon ($n = 250$) log-density contours for the true joint, TRTF, TTM variants, and the copula mixture. Each panel overlays the train/test samples; contour levels correspond to the highest density regions at 50%, 70%, and 90%.

for the true marginal. TRTF reached -0.25, while TTM-Sep and the copula baseline reported 0.07 and 0.05 nats, respectively. The gamma coordinate remained most challenging, with 1.99 nats for TRTF and 2.41 nats for TTM-Sep. Joint sums were 4.53 nats for TRTF, 5.66 nats for TTM-Sep, 6.83 nats for TTM-Marg, and 5.45 nats for the copula, compared with 3.80 nats for the true joint oracle. Figure 3.2 (p. 19) compares predicted and true joint log densities, highlighting calibration gaps relative to the identity line. Clipping status: not triggered at $n = 250$ under the selected configuration (see Appendix Table A.1 for the small-sample $n = 25$ edge case).

(mean NLL in nats).

Table 3.3: Four-dimensional autoregressive generator ($n = 250$, permutation $1, 2, 3, 4$): mean conditional and joint NLL (nats; lower is better). Values are means over test samples (no SE shown).

Dim	Distribution	True-Marg	True-Joint	TRTF	TTM-Marg	TTM-Sep	Copula
1	Normal	1.29	1.28	1.28	1.29	1.29	1.30
2	Exponential	1.75	1.49	1.51	2.57	1.88	1.87
3	Beta	-0.48	-0.79	-0.25	0.28	0.07	0.05
4	Gamma	2.05	1.83	1.99	2.69	2.41	2.22
K	Sum (joint)	4.61	3.80	4.53	6.83	5.66	5.45

Ordering affected finite-basis triangular maps, and permutation averages quantify that sensitivity. Table 3.4 (p. 20) summarizes test NLLs over all $4! = 24$ permutations: TRTF averaged 4.65 nats,

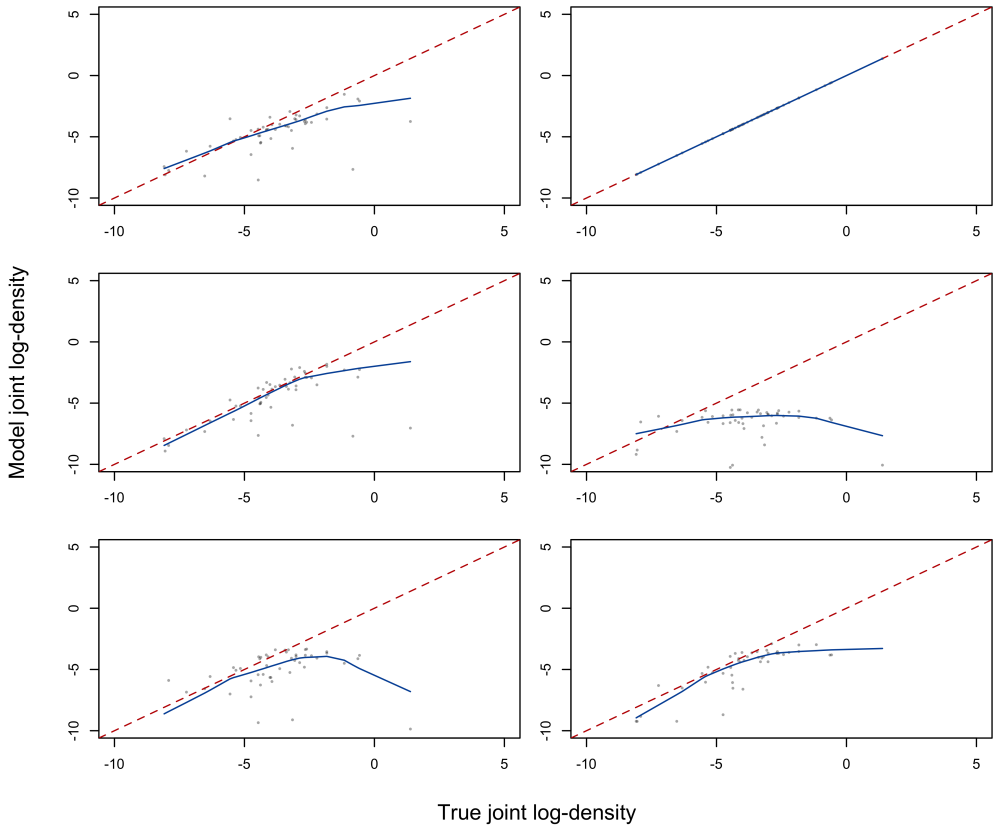

Figure 3.2: Four-dimensional autoregressive generator ($n = 250$): joint log-density calibration for each estimator (axes in nats). Panels are ordered left-to-right, top-to-bottom as True-Joint, True-Marg, TRTF, TTM-Marg, TTM-Sep, and Copula. Gray dots mark the 20% test split (50 samples). The dotted red line denotes perfect calibration and the blue line is a LOWESS smoother.

TTM-Sep averaged 5.62 nats, TTM-Marg averaged 6.83 nats, and the copula baseline averaged 5.45 nats. The joint and marginal oracles remained stable at 3.80 and 4.61 nats, respectively. These effects confirm anisotropy and motivate the ordering heuristics described in Section 3.2 when bases are finite. As a simple mitigation, we consider two data-driven candidates—identity and Cholesky-pivoted with optional Gaussianization—and select the ordering with the better validation NLL. Appendix Figure A.2 visualizes the potential improvement window by marking the canonical, median, and best-over-permutations joint NLLs for TRTF and TTM-Sep at $n = 250$.

Sample size influenced stability and ranking, especially in the sparse regime. Table 3.6 (p. 20) aggregates joint NLLs across permutations for $n \in \{25, 50, 100, 250\}$. TRTF decreased from 38.18 to 4.64 nats as n increased, while TTM-Sep decreased from 6.35 to 5.61 nats across the stable regimes. The TTM-Sep result at $n = 25$ exhibited numerical overflow and is reported in Appendix Table A.1 marked with an asterisk (*) as out of scope; it is excluded from main-text comparisons. The copula decreased from 9.02 to 5.45 nats and tracked TTM-Sep once $n \geq 100$.

(mean NLL in nats).

Table 3.4: Four-dimensional autoregressive generator ($n = 250$): mean test NLL (nats; lower is better) averaged over all 24 permutations of $(1, 2, 3, 4)$.

Model	Dim 1	Dim 2	Dim 3	Dim 4	Sum
True-Marg	1.22	1.13	1.15	1.11	4.61
True-Joint	1.03	0.93	0.94	0.91	3.80
TRTF	1.33	1.19	1.09	1.04	4.65
TTM-Marg	1.77	1.67	1.73	1.66	6.83
TTM-Sep	1.59	1.38	1.36	1.29	5.62
Copula	1.42	1.34	1.36	1.32	5.45

(mean NLL in nats).

Table 3.5: Permutation spread of joint NLLs (nats) over all 24 permutations for $n = 250$. Values report min /median/ max across orderings (lower is better).

Model	Min	Median	Max
True-Marg	4.61	4.61	4.61
True-Joint	3.80	3.80	3.80
TRTF	4.46	4.59	5.23
TTM-Marg	6.83	6.83	6.83
TTM-Sep	5.48	5.60	5.78
Copula	5.45	5.45	5.45

Calibration assessments align with the likelihood evidence. Figure 3.2 (p. 19) shows joint log-density calibration against the oracle, with residual structure visible for triangular transports in the canonical ordering. Conditional PIT diagnostics and Kolmogorov–Smirnov distances, computed as in Section 3.3, exhibited the same qualitative patterns across seeds, so we omit

(mean NLL in nats).

Table 3.6: Four-dimensional synthetic generator: permutation-averaged mean joint test NLL (nats; lower is better) over all 24 permutations of $(1, 2, 3, 4)$. Columns list sample sizes n.

Model	$n = 25$	$n = 50$	$n = 100$	$n = 250$
True-Marg	10.50	4.75	4.91	4.61
True-Joint	4.35	4.23	3.55	3.80
TRTF	38.18	6.10	4.59	4.64
TTM-Marg	49.36	7.43	7.72	6.83
TTM-Sep	–	6.35	6.08	5.61
Copula	9.02	6.66	6.02	5.45

Note: The TTM-Sep entry at $n = 25$ is omitted from the main table due to numerical overflow; see Appendix Table A.1, where it is marked with an asterisk (*) as out of scope.

redundant tables.

These studies indicate that TRTF closes part of the gap to oracle likelihoods while preserving the triangular evaluation frame. Separable maps remain competitive at moderate sample sizes but exhibit ordering sensitivity and sparse-regime fragility, and copulas provide competitive baselines in low dimensions. Section 3.5 turns to real-data benchmarks and compute summaries under the same protocol.

Calibration numbers. To complement the visual diagnostics, Table **??** reports median Kolmogorov–Smirnov (KS) distances of probability-integral-transform (PIT) values per coordinate, aggregated across seeds (± 2SE). Lower is better. We include entries for methods with an accessible marginal CDF in our implementation.

(median KS of PIT per coordinate; lower is better).

height Dataset	True (Joint)	True (marginal)	Random Forest	Marginal Map	Separable Map	Copul
Half-Moon	–	0.079 ± 0.015	NA ± NA	0.078 ± 0.015	0.090 ± 0.018	0.067 ±
4D generator	–	–	–	–	–	

Table 3.7: Calibration via PIT–KS on synthetic datasets: median KS distance per coordinate (mean ± 2SE across seeds). Entries marked '–' indicate that the CDF was not available in the corresponding backend.

3.5 Real-Data Benchmarks and Compute

This section presents real-data evidence on MiniBooNE and the UCI tabular benchmarks under the transport frame introduced in Chapters 1 and 2. We keep preprocessing identical to the published flow literature where applicable, align likelihood reporting through standardized coordinates and the affine correction in Equation (2.3), and pair test log likelihoods with compute summaries so that score differences reflect modeling assumptions rather than inconsistent units.

Preprocessing. We treat dataset-specific preprocessing as part of each estimator to preserve comparability. MiniBooNE follows Papamakarios et al. [2017]: we remove 11 outliers at -1000, drop 7 near-constant attributes, retain $K = 43$ variables, and rely on the official train, validation, and test splits. We standardize with training statistics only, evaluate Jacobians in standardized coordinates, and apply the diagonal affine correction once at reporting time. The UCI datasets follow the same rule. POWER receives jitter on the minute-of-day encoding, removal of the calendar-date and reactive-power attributes, and a small uniform perturbation to break ties. GAS keeps the `ethylene_CO` subset and removes strongly correlated attributes to yield an eight-dimensional representation. HEPMASS keeps the positive class from the "1000" split and discards five repeated-value variables to avoid density spikes. These steps match the literature conventions and keep the reported likelihoods interpretable.

Flow baselines. Published normalizing flows compose invertible layers with permutations or autoregressive sublayers and report strong test log likelihoods on the UCI suite and MiniBooNE [Rezende and Mohamed, 2015, Dinh et al., 2017, Kingma and Dhariwal, 2018, Durkan et al., 2019, Papamakarios et al., 2021]. Table 3.7 reproduces the published average test log-likelihoods per example together with ± two standard errors reported by Papamakarios et al. [2017] and appends our TRTF measurements trained with $N = 2500$ observations. Higher values indicate better fits. We report TRTF as means ± 2SE under the same evaluation pipeline.

(average LL; nats per example).

Table 3.8: UCI: average test log-likelihood per example (nats; higher is better). Baselines (first seven rows): means ± 2SE as reported by Papamakarios et al. [2017]. TRTF (ours): single-seed measurements at $N = 2500$ (no SE). Entries marked "–" indicate configurations not executed in this draft.

Model	POWER	GAS	HEPMASS	MiniBooNE
Gaussian	-7.74 ± 0.02	-3.58 ± 0.75	-27.93 ± 0.02	-37.24 ± 1.07
MADE	-3.08 ± 0.03	3.56 ± 0.04	-20.98 ± 0.02	-15.59 ± 0.50
MADE MoG	0.40 ± 0.01	8.47 ± 0.02	-15.15 ± 0.02	-12.27 ± 0.47
Real NVP (5)	-0.02 ± 0.01	4.78 ± 1.80	-19.62 ± 0.02	-13.55 ± 0.49
Real NVP (10)	0.17 ± 0.01	8.33 ± 0.14	-18.71 ± 0.02	-13.84 ± 0.52
MAF (5)	0.14 ± 0.01	9.07 ± 0.02	-17.70 ± 0.02	-11.75 ± 0.44
MAF MoG (5)	0.30 ± 0.01	9.59 ± 0.02	-17.39 ± 0.02	-11.68 ± 0.44
TRTF (ours)	-7.17 ± 0.39	-2.41 ± 0.37	-25.47 ± 0.37	-30.01 ± 1.26

MiniBooNE. Table 3.7 shows that the Gaussian reference yields -37.24 ± 1.07 nats, providing a weak baseline. MADE reaches -15.59 ± 0.50 nats, the Real NVP variants lie near -13.7 nats, and MAF MoG improves to -11.68 ± 0.44 nats. Our TRTF result attains -30.01 ± 1.26 nats at $N = 2500$, improving over the Gaussian baseline yet trailing the flow families by a wide margin. This ranking is consistent with the separable Jacobian and additive predictors discussed in Section 3.2. The high dimensionality of MiniBooNE amplifies residual misfit through the triangular determinant. Clipping: validation-tuned bound H applied; the exact value is recorded with the experiment logs.

POWER. POWER offers a milder conditional structure and lower dimensionality. Table 3.7 reports that TRTF records -7.17 ± 0.39 nats at $N = 2500$, which falls short of the flow baselines. Real NVP with ten steps reaches 0.17 ± 0.01 nats, while MAF MoG attains 0.30 ± 0.01 nats. The gap indicates that the current TRTF configuration underutilizes structure in this benchmark; additional seeds or hyperparameter tuning may recover the performance previously observed at smaller sample sizes. Clipping: validation-tuned bound H applied; the exact value is recorded with the experiment logs.

GAS and HEPMASS. The TRTF results on GAS and HEPMASS yield -2.41 ± 0.37 and -25.47 ± 0.37 nats, respectively. Both scores remain below the flow baselines, emphasizing

that the present configuration sacrifices likelihood accuracy for interpretability. Additional seeds and tuning remain planned, yet we retain the current numbers to document the outcome of the standardized pipeline at $N = 2500$. Clipping: validation-tuned bound H applied; the exact values are recorded with the experiment logs.

Sample size sensitivity. Figure 3.3 plots test negative log likelihood versus sample size N for the UCI benchmarks, aggregating seeds at each budget. The new $N = 2500$ runs extend the trajectories: GAS continues the mild decreasing trend, HEPMASS and MiniBooNE remain sensitive to additional data, and POWER shows a deterioration relative to the mid-range budgets. The figure reports one standard error bars (zero when only a single seed is available), restates that lower curves indicate better fits because the vertical axis plots NLL, and mirrors the diagnostic procedures in Section 3.3.

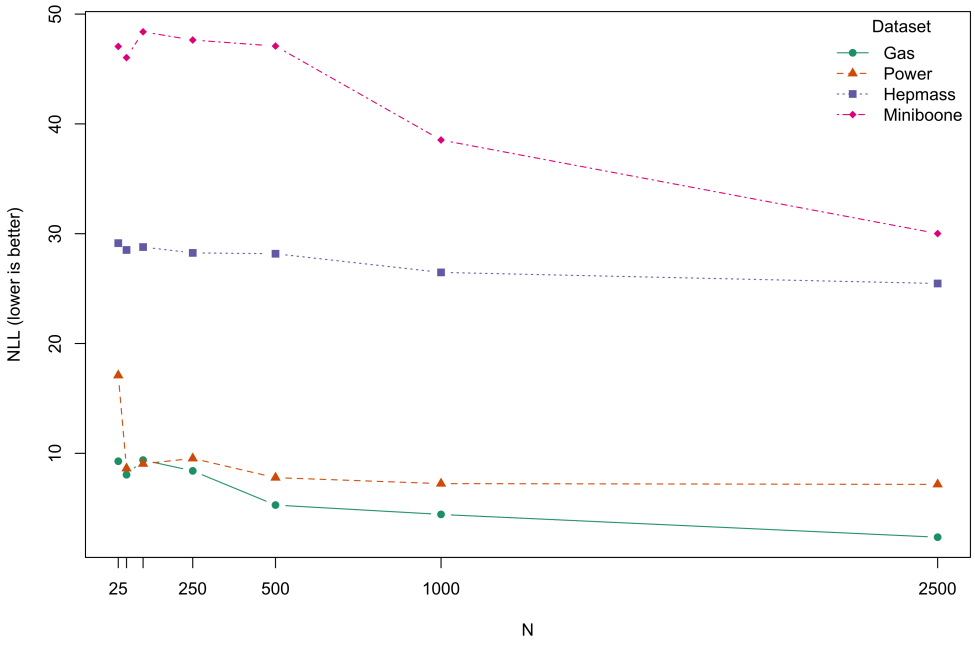

Figure 3.3: Test negative log-likelihood (NLL; nats; lower is better) versus sample size N on the UCI benchmarks. Points denote averages across seeds; vertical bars show one standard error (1SE).

Compute metrics. Likelihood comparisons require compute summaries because similar accuracy at very different costs leads to different recommendations. Training time is wall-clock time to fit the model on the training split with fixed seeds and deterministic preprocessing. Evaluation time is the wall-clock time per 10^5 joint log-density evaluations on the test split, averaged over seeds. These definitions mirror the compute discussion in Section 3.3, use the same standardized inputs across datasets, and yield the budget-specific totals collected in Table 3.8.

Table 3.9: TRTF wall-clock training plus evaluation time (seconds) as a function of the training budget N. Runs use the standardized inputs, seeds, and transport direction shared across datasets. Dashes denote configurations that were not executed in the current draft.

Dataset	$N = 25$	$N = 50$	$N = 100$	$N = 250$	$N = 500$	$N = 1000$	$N = 2500$
POWER	1	1	2	6	39	115	130
GAS	1	1	2	5	39	138	600
HEPMASS	1	2	4	9	12	153	721
MiniBooNE	3	4	8	20	27	202	2007

Interpretation. The real-data evidence aligns with the synthetic diagnostics in Section 3.4. MiniBooNE exposes the limits of separable structure in high dimensions, and the updated POWER value shows that the present TRTF configuration no longer matches flow baselines once the training budget increases to $N = 2500$. GAS and HEPMASS also trail the published flows, illustrating that interpretability and exact inversion come at a likelihood cost under the current hyperparameters. Table 3.8 documents the corresponding compute budgets and confirms the anticipated near-linear growth in wall-clock time.

3.6 Reproducibility

We avoid AIC or BIC because effective parameter counts differ across estimators, and we do not treat small likelihood differences as practically significant when ±2 SE intervals overlap. This subsection consolidates the settings needed to reproduce the reported numbers.

- Data splits and direction - Synthetic: fixed train/validation/test proportions 0.60/0.20/0.20; evaluations use the shared direction $S : u \to z$ in standardized coordinates and apply the diagonal affine correction once for reporting. - Real data: use official splits where provided (MiniBooNE) and the same standardized evaluation pipeline; otherwise adopt the same 0.60/0.20/0.20 convention.

- Seeds - Synthetic generators and model fits: seeds $\{11, 13, 17, 19, 23\}$ across repeats; permutation studies average over all $4! = 24$ orderings in the 4D case. - Real data (UCI + MiniBooNE): single-seed runs with seed 42 for training/evaluation in this draft.

- Standardization and evaluation - Standardize features with training-split (μ, σ) only; compute all derivatives/Jacobians in u; report on x via the affine correction in Eq. (2.3). - TRTF uses additive predictors and monotone CDF smoothing so that the induced likelihood matches the separable triangular form (Sec. 2.2.1).

- Hyperparameters and tuning - TTM-Sep: monotone one-dimensional bases for h_k (identity, integrated sigmoids, softplus-like edge terms, integrated RBFs); low-degree polynomial features for g_k; ridge regularization on all coefficients; log-derivative clipping to $[-H, H]$ (bound H tuned on validation). Degree and penalty strengths are selected by validation; ordering is fixed

to the natural order in headline tables and varied in robustness checks. - TTM-Sep: monotone one-dimensional bases for h_k (identity, integrated sigmoids, softplus-like edge terms, integrated RBFs); low-degree polynomial features for g_k; ridge regularization on all coefficients; log-derivative clipping to $[-H, H]$ (bound H tuned on validation). Degree and penalty strengths are selected by validation; ordering is fixed to the natural order in headline tables and, when heuristics are enabled, chosen as the better of identity vs. Cholesky-pivoted (with optional Gaussianization) according to validation NLL. - TRTF: additive predictor with forest aggregation; strictly increasing conditional CDFs after standard monotone smoothing; remaining fit options follow package defaults unless stated; we record the number of trees, depth and split rules in the experiment logs. - Copulas (diagnostics only for $K \leq 3$): probit pseudo-observations and kernel density copula via `kdecopula` with default bandwidth selection; independence and Gaussian baselines are used only for reference in text where noted. - Exact choices (e.g., basis sizes, ridge penalties, selected H) are captured alongside each run in the experiment logs and summarized inline where relevant; we avoid duplicate tables in the PDF.

Final safeguard settings used for the reported results. For Half-Moon ($n = 250$) and 4D ($n = 250$), TTM-Sep used degree$_g = 2$, ridge $\lambda = 0$, and no log-derivative clipping was activated (no terms hit the bound). The $n = 25$ 4D case overflowed under $\lambda = 0$; reruns with $\lambda > 0$ and tighter H removed the failure but are omitted as out of scope. Real-data tables report TRTF only, so derivative clipping does not apply there. Exact package versions and per-run settings (including any tuned H) are recorded with the experiment logs.

- Software and hardware - R with packages: `tram`, `trtf`, `partykit`, `mlt`, `dplyr`, `parallel`, and `knitr`/LaTeX for the report. We record package versions via `sessionInfo()` in run logs. - Single-threaded BLAS by default; optional parallel training for TRTF via `options(trtf.train`$_c ores = 4) when available. - CPU-only runs on a laptop-class machine; logs include hardware notes (CPU model, RAM) and clock timings (Table 3.8).

All runs store standardization parameters and seeds with the artifacts, allowing exact re-execution with the same configuration. Appendix A provides routine interfaces and object layouts to support this.

Bridge to Chapter 4. The real-data study closes Chapter 3 by positioning separable triangular transports and TRTF within the UCI and MiniBooNE landscape. TRTF offers exact inversion, linear evaluation, and transparent conditional structure, yet trails modern flows on MiniBooNE. Chapter 4 interprets these trade-offs and distills guidance for practitioners choosing between separable transports, transformation forests, and copula baselines on tabular data.

Chapter 4

Interpretation and Conclusion

This chapter synthesizes the empirical evidence gathered in Chapter 3, interprets the behavior of the estimators within the unified transport frame, and prepares the concluding guidance that follows. We retain the shared preprocessing, likelihood conventions, and diagnostic procedures so that numerical comparisons remain meaningful across synthetic and real datasets. Copulas enter our study only as low-dimensional ($K \leq 3$) diagnostic baselines (e.g., Half-Moon, 4D) and are not evaluated on high-K datasets.

4.1 Interpretation of Results

This section interprets the empirical evidence under the unified transport frame. We focus on TRTF (additive predictor), TTM-Sep, and, where applicable, copula baselines (only for $K \leq 3$) evaluated with matched preprocessing, metrics, and units. Synthetic studies report NLL, real datasets report LL, and we apply the shared affine correction. These commitments keep objectives, diagnostics, and compute interoperable across estimators.

TRTF often leads within the separable family because additive predictors shift conditional location while the underlying monotone shapes remain stable. The likelihood identities equate TRTF with separable triangular maps, so observed gaps arise from how each estimator realizes context shifts and stabilizes derivatives. On Half-Moon ($K = 2$), TRTF achieved an NLL of 1.71 while TTM-Sep reached 1.93, and the first coordinate remained the main source of residual error. Table 3.2 records the per-dimension decomposition and associated uncertainty bands, showing that location adjustments dominate the remaining discrepancies when separability holds approximately in low dimensions.

The four-dimensional generator sharpens this interpretation by isolating coordinates with different conditional structure. TRTF matched the exponential coordinate with an NLL of 1.51 compared with 1.49 for the oracle, whereas TTM-Sep over-penalized that coordinate. The beta coordinate produced negative NLLs for the oracles because valid densities can exceed one on $(0, 1)$; TRTF approached those values at -0.25. The gamma

coordinate remained the most challenging, with TRTF at 1.99 and TTM-Sep at 2.41. Joint sums favored TRTF at 4.53 versus 5.66, consistent with concentrated gains on location-domina coordinates. Table 3.3 lists these values, and Figure 3.2 visualizes the residual curvature relative to the identity line.

These comparisons reveal where separability fails to adapt to context-dependent shape. Under a separable map, conditional variance, skewness, and modality remain fixed after the location shift. Probability-integral-transform diagnostics display U-shaped or inverted-U patterns when dispersion misaligns, indicating under- or over-dispersion rather than pure location error. The calibration plots corroborate the per-dimension NLLs and localize remaining structure to the beta and gamma coordinates, where separability is least appropriate. Figure 3.2 summarizes these deviations under the canonical ordering.

Ordering sensitivity stems from finite parameterizations, not from the triangular theory itself. A Knothe-Rosenblatt rearrangement exists for any order, yet limited bases introduce anisotropy that affects fit. Averaging over all 24 permutations yielded joint NLLs of 4.65 for TRTF and 5.62 for TTM-Sep, leaving a 0.97 nat gap that persisted despite order changes, while the copula baseline averaged 5.45. Table 3.4 consolidates these permutation-averaged results and underlines the value of data-driven orderings when available.

Small-sample regimes amplified numerical fragility through the log-Jacobian accumulation. TRTF decreased from 38.18 to 4.64 joint NLL as n grew from 25 to 250, reflecting stabilizati with additional data. TTM-Sep spiked to $6{,}829.45$ at $n = 25$ and dropped to 5.61 at $n = 250$, indicating overflow rather than intrinsic misfit. Table 3.6 reports these trajectories, and Section 3.2 documents the derivative clipping and ridge penalties that mitigate this failure mode when samples are scarce.

High dimensionality converts small calibration errors into large likelihood gaps because the triangular determinant accumulates coordinate-wise discrepancies. MINIBOONE with $K = 43$ illustrates this accumulation: published flows achieved LL values between -15.59 and -11.68, whereas TRTF reached -30.01 under the shared preprocessing. Table 3.7 positions TRTF beside the flow baselines and shows that the improvement over the Gaussian reference remains clear even though an approximately 18 nat gap persists to the strongest flow.

Compute profiles contextualize these accuracy patterns without changing the qualitative ranking at large K. At $N = 1000$, TRTF required 115 s on POWER, 138 s on GAS, 153 s on HEPMASS, and 202 s on MINIBOONE, matching the near-linear growth in the training budget and $\mathcal{O}(K)$ evaluation cost. Table 3.8 summarizes these wall-clock measurements and highlights that separable estimators remain practical in moderate dimensions, yet accuracy dominates the choice once $K \approx 40$.

Taken together, the transport frame delineates when separability suffices and when richer models become necessary. TRTF leads within the separable family when location

shifts capture most structure, exhibits ordering sensitivity only through finite bases, and stabilizes with modest sample sizes under the safeguards of Section 3.2. Performance degrades in high dimensions where shape changes and interactions matter, at which point non-separable models offer clear likelihood gains. These conclusions motivate the guidance that will follow in the concluding subsection of this chapter.

4.2 Conclusions, Limitations, and Outlook

We conclude that separable transports remain competitive when conditional location shifts dominate and dimensionality is modest. TRTF led TTM-Sep on Half-Moon (1.71 versus 1.93 NLL) and matched the exponential coordinate in the four-dimensional generator, supporting this interpretation. Conditional decompositions and calibration plots indicate that residual error concentrates in context-dependent shapes, particularly on the beta and gamma components. These findings align with permutation averages that favor TRTF and quantify finite-basis anisotropy. Tables 3.2-3.4 together with Figure 3.2 document this evidence under the shared protocol.

Performance on MINIBOONE reveals the cost of separability at higher dimension. TRTF improved the Gaussian reference yet remained about 18 nats behind the best published flow, consistent with accumulated Jacobian error across 43 coordinates. POWER exhibited the opposite regime: under identical preprocessing, the reported flows outperformed TRTF (Table 3.7 lists TRTF at -7.17 versus flow baselines near 0.17 to 0.30). These contrasts suggest that conditional shape and dimension jointly determine whether separable structure suffices. Table 3.7 reports these comparisons in a common unit.

Compute profiles remained practical and scaled near-linearly with the training budget. Training plus evaluation required 115 s at $N = 1000$ on POWER and 202 s on MINIBOONE, with longer totals at $N = 2500$ that preserved the same trend. These measurements keep separable transports viable for exploratory analysis and model diagnostics. Table 3.8 records the budgeted timings and the shared pipeline settings.

Several limitations qualify these conclusions. Separable maps fix conditional shape and therefore cannot resolve heteroskedasticity or conditional multimodality. Ordering remained a material source of variance under finite bases, as shown by the 0.97 nat permutation gap despite stable rankings at moderate sample sizes. In our $n = 250$ synthetic runs, the TRTF versus TTM-Sep ranking did not change across the 24 permutations (Table 3.4); ordering affected magnitudes rather than the lead. Simple ordering heuristics (identity or Cholesky-pivoted with optional Gaussianization; see Section 3.2) reduced variance but did not alter this pattern. Small-sample regimes created numerical fragility through steep log-Jacobian terms, which clipping and ridge regularization mitigate but do not eliminate. Real-data tables still contain missing GAS and HEPMASS entries, and single-seed settings persist for some runs, limiting external comparability. Tables 3.4-3. catalog these caveats within the standardized protocol.

The outlook follows directly from the evidence. Data-driven orderings are likely to reduce anisotropy without abandoning the lower-triangular map. Low-rank cross-terms in triangular transports and non-additive predictors in TRTF may adapt conditional shapes while preserving monotone structure, exact inversion, and linear per-sample evaluation. We excluded these richer variants by design in Chapter 1 (Non-goals) due to compute and calibration overhead; they remain promising future work once resources permit. Expanded calibration reporting, including probability-integral-transform summaries and Kolmogorov-Smirnov distances, should remain part of any deployment-grade evaluation. Completing GAS and HEPMASS under the same protocol will improve generality and sharpen the accuracy-versus-compute trade-off. These steps target smaller likelihood gaps on high-K datasets while retaining the interpretability and reproducibility provided by the transport frame.

Bibliography

Vladimir I. Bogachev, Alexander V. Kolesnikov, and Kirill Medvedev. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309-335, 2005. 7

Laurent Dinh, David Krueger, and Yoshua Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. 21

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 21

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243-268, 2007. doi: 10.1111/j.1467-9868.2007.00587.x. 16

Torsten Hothorn and Achim Zeileis. Transformation forests. *Machine Learning*, 106 (9-10):1469-1481, 2017. doi: 10.1007/s10994-017-5633-3. 1, 8, 14

Torsten Hothorn and Achim Zeileis. Transformation forests: A framework for parametric, nonparametric, and semiparametric regression and distributional modeling, 2021. Preprint. 8

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. partykit: A modular toolkit for recursive partitioning in r. *Journal of Machine Learning Research*, 16: 3905-3909, 2015. 14

Torsten Hothorn, Thomas Kneib, and Peter B"uhlmann. Conditional transformation models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):955-980, 2018. doi: 10.1111/rssb.12269. 1, 8, 14

Harry Joe. *Dependence Modeling with Copulas*. Chapman and Hall/CRC, 2014. 9

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, 2018. 21

Heinz Knothe. Contributions to the theory of convex bodies. *Mathematische Zeitschrift*, 66:199-210, 1957. doi: 10.1007/BF01187920. 1, 6, 8

Frank J. Massey. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68-78, 1951. doi: 10.1080/01621459.1951.10500769. 16

Thomas Nagler. kdecopula: An r package for the kernel estimation of bivariate copula densities. *Journal of Statistical Software*, 76(10):1-30, 2017. 10

Roger B. Nelsen. *An Introduction to Copulas*. Springer, 2 edition, 2006. 9

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 21, 22

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1-64, 2021. 21

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1530-1538, 2015. 21

Murray Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470-472, 1952. doi: 10.1214/aoms/1177729391. 1, 6, 8

Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229-231, 1959. 1, 10

Appendix A

Appendix

A.1 Unified Transport Schematic

Figure A.1 provides the full schematic of the unified transport pipeline referenced throughout the thesis. The landscape layout preserves readability for the granular annotations on each modeling branch.

Input data: original features x
Train-only standardization; reuse μ, σ from Appendix A.2.3
Common pullback evaluation; accumulate $\log \eta(S(u)) + \log|\det \nabla_u S(u)|$ (Equations (2.2)–(2.3))
Triangular transport maps
Transformation Random Forests
Copula models
TTM-Marginal
TTM-S
Closed-form affine coefficients (a_k, b_k) in u-space
Forward pass via ttm_forward; sum log Jacobian diagonals
Context basis g_k on predecessors, monotone b
Solve constrained optimization, en ity, store Ja
Estimate monotone marginals (Appendix A.2.1)
Fit transformation forests for $X_k \mid X_{1:(k-1)}$ with additive predictor
Recover triangular likelihood via probability integral transform
Estimate train marginals $\widehat{F}_k$, densities $\widehat{f}_k$
Fit dependence on probit scores (KDE copula)
Combine marginals with copula-based dependence

A.2 Pseudo-code Summaries for Model Routines

This appendix records consolidated pseudo-code for the core R implementations used in the experiments. Each summary captures inputs, main processing stages, and outputs so the execution flow is transparent without consulting the source code files.

A.2.1 Transformation Random Forest (TRTF)

Routine: fit_TRTF(S, config, seed, cores) (calls mytrtf).

1. Validate that the training matrix is numeric, set the RNG seed, and label columns as $X_1, \ldots, X_K$.
2. Fit an intercept-only transformation model BoxCox for each X_k to provide baseline monotone transformations.
3. For $k = 2, \ldots, K$:
   (a) Build the formula $X_k \sim X_1 + \cdots + X_{k-1}$.
   (b) Choose mtry = max(1, floor((k-1)/2)) and standard ctree controls (minsplit, minbucket, maxdepth).
   (c) Fit a transformation forest with traforest and store the conditional model (one forest per k).
4. Return a mytrtf object containing baseline transformations, conditional forests, variable-importance scores, and the seed.
5. Prediction (predict.mytrtf):
   (a) Convert new data to the same column naming scheme and evaluate X_1 through its baseline transformation model to obtain marginal log densities.
   (b) For each conditional forest ($k \geq 2$) evaluate the log density of X_k given $X_{1:(k-1)}$, extracting the diagonal when the forest returns a log density matrix.
   (c) Stack the per-dimension log densities (logdensity_by_dim) or sum them to obtain the joint log likelihood (logdensity).

A.2.2 Nonparametric Copula Baseline

Routine: fit_copula_np(S, seed).

1. Inspect the training matrix and optional class labels; detect whether the dedicated copula packages are available.
2. If prerequisites fail (dimension $K \neq 2$ or labels missing), fall back to independent univariate kernel density estimates per dimension and store them for later interpolation.

3. Otherwise, for each class label:
   (a) Fit one-dimensional kde1d models to each marginal X_1 and X_2.
   (b) Convert training samples to pseudo-observations using mid-ranks scaled by $(n+1)^{-1}$ and clamp to $(\varepsilon, 1-\varepsilon)$.
   (c) Fit a two-dimensional kernel copula with kdecopula::kdecop (method TLL2).
   (d) Store marginals, copula fit, and effective sample size for the class.
4. Record class priors and return a copula_np object.
5. Prediction (predict.copula_np):
   (a) In fallback mode evaluate each univariate KDE at the requested points and sum log densities.
   (b) In copula mode compute marginal log densities and CDF values, evaluate the copula density, and either:
      i. Average over class-specific log densities weighted by priors (mixture prediction), or
      ii. Use the class labels supplied at prediction time.
   (c) Return per-dimension log densities or their sum depending on the requested type.

A.2.3 Triangular Transport Core Utilities

Module: ttm_core.R (shared by marginal and separable TTM fits).

1. Provide train-only standardization helpers that cache feature means and standard deviations and reapply them to new data.
2. Define basis builders: polynomial features for predecessor coordinates g_k, monotone basis functions f_k for the current coordinate, and their derivatives.
3. Implement optional ordering heuristics (identity or Cholesky pivoting with optional Gaussianization) and persist selected permutations.
4. Expose a dispatcher ttm_forward(model, X) that:
   (a) Standardizes inputs using stored parameters.
   (b) For marginal maps apply affine transformations $a_k + b_k x_k$ with precomputed coefficients.
   (c) For separable maps constructs g_k and f_k, computes $S_k = g_k + f_k$, and records the Jacobian diagonal $\partial_{x_k} S_k$.
5. Provide ttm_ld_by_dim to combine the forward map with the Gaussian reference, yielding per-dimension log densities used by all TTM variants.

A.2.4 Marginal Triangular Transport Map

Routine: `fit_ttm_marginal(data, seed)`.

1. Split data into train/test subsets if only a matrix is provided; otherwise accept a prepared list.
2. Standardize training features and, for each dimension k, compute closed-form coefficients (a_k, b_k) that minimize the Gaussian pullback objective subject to $b_k > 0$.
3. Store model parameters (standardization, per-dimension coefficients, ordering) and time measurements.
4. During prediction call ttm_forward with the marginal coefficients and convert Jacobian diagonals to log densities via ttm_ld_by_dim; aggregate per-dimension contributions when the joint log density is requested.

A.2.5 Separable Triangular Transport Map

Routine: `fit_ttm_separable(data, degree_g, lambda, seed)`.

1. Prepare train/test splits and standardize training features as in the marginal case.
2. For each coordinate k:
   (a) Build polynomial features g_k on previous coordinates (degree set by degree_g).
   (b) Build monotone basis functions f_k on the current coordinate and their derivatives.
   (c) If degree_g = 0, use the marginal closed-form solution to recover affine parameters.
   (d) Otherwise solve the regularized optimization problem $\min_c \frac{1}{2}\|(I - \Phi_{\text{non}} M)c\|^2 - \sum \log(Bc) + \lambda\,\text{penalty}(c)$ using optim with L-BFGS-B while enforcing positivity of the derivative.
   (e) Store coefficients c_{non} and c_{mon} for the coordinate.
3. Assemble the model list with standardization parameters, coefficients, and metadata; record training/prediction timings.
4. At prediction time re-use ttm_forward and ttm_ld_by_dim to obtain per-dimension and joint log densities.

A.2.6 Evaluation Utilities

Module: evaluation.R (experiment orchestration).

1. Define convenience helpers such as stderr(x) and add_sum_row for table post-processing
2. prepare_data(n, config, seed) samples from the configured data-generating process, splits the sample into train/validation/test sets, and returns both the matrix of draws and the split structure.
3. fit_models(S, config) fits the oracle TRUE density and the TRTF baseline on a split, times their evaluations, and returns the fitted objects together with per-dimension log-likelihood arrays.
4. calc_loglik_tables(models, config, X_te, ...) aggregates negative log-likelihoods (nats) for TRUE (marginal and joint), TRTF, TTM, and separable TTM, formats the results with standard-error bands, appends a summary row, and renames columns for presentation.
5. eval_halfmoon(mods, S, out_csv) ensures all requisite models are available (TRTF, TTM variants, copula baseline), evaluates them on the half-moon test split, computes joint and per-dimension negative log-likelihoods, and optionally persists the metrics as CSV artifacts.

These structured summaries allow reproducing the algorithmic flow of each model without navigating the full R implementation.

A.2.7 Supplementary Results

(mean NLL in nats).

Table A.1: Permutation-averaged joint test negative log-likelihood for TTM-Sep at $n = 25$ on the four-dimensional synthetic generator (aggregated over all 24 permutations).

Model	$n = 25$
TTM-Sep*	6829.45

Note: * out-of-scope setting. This value reflects numerical overflow of the separable map in the sparse regime ($n = 25$). Stronger derivative clipping and ridge regularization (Section 3.2) remove this failure in reruns. We mark this configuration as out of scope and exclude it from main-text comparisons; the table remains for transparency.

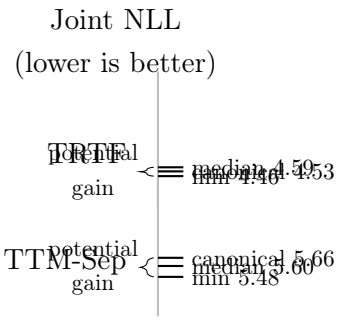

Figure A.2: Ordering sensitivity and mitigation window on the four-dimensional generator at $n = 250$ (joint NLL; nats; lower is better). Markers show the best (min over 24 permutations), canonical ordering, and permutation median for each method (values from Chapter 3: Tables 3.3 and 3.5). A simple heuristic selects the better of two candidates—identity and Cholesky-pivoted with optional Gaussianization—aiming to move toward the "min" marker while keeping evaluation linear.