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

---

Master Thesis in Biostatistics (STA495)

by

LÃ©on Kia Faro
13-795-026

supervised by

Prof. Dr. Torsten Hothorn
Name of other supervisor (with title and affiliation if external)
Name of assistant (if applicable)
Name of responsible faculty member (if different from first line)

Zurich, September 2025

Abstract

Acknowledgments

Contents

Chapter 1

Introduction

Estimating the joint density $\pi_X(x)$ of a random vector $x \in \mathbb{R}^K$ underpins probabilistic modelling, anomaly detection, simulation-based inference, and decision making under uncertainty. Modern applications combine high dimension with structure that changes across contexts: conditional variance may grow with predictors, skewness can flip sign across regions, and mixtures can create multimodal slices. A coherent response is *measure transport*: we couple the unknown target with a simple reference through an invertible transformation and exploit that coupling to compute likelihoods, draw samples, and evaluate conditionals. Throughout we adopt the notation established in a recent tutorial; Chapter 2 summarises the relevant background. We standardise observations, learn monotone triangular maps to a Gaussian reference, and report log densities on the original scale using a fixed affine correction. All Jacobians and partial derivatives involving the transport are taken in the standardised space unless stated otherwise.

This thesis compares three model classes that realise the transport programme with different inductive biases: triangular transport maps parameterised directly, transformation random forests that induce transports via the probability integral transform, and copulas that decouple marginals from dependence. Each offers a distinct view on conditional density estimation and raises the question of when separable structure suffices and when richer cross-term capacity is required. The following sections outline our contributions, the research questions we address, and the scope of the work.

1.1 Contributions and research questions

We make three contributions within a single, notation-coherent framework.

1. *Theory.* We clarify the link between transformation forests and triangular transport and formalise when the induced maps are separable (Hothorn and Zeileis, 2017, 2021; Hothorn *et al.*, 2018).

2. *Empirics.* We compare separable and cross-term triangular maps, transformation forests under additive predictors, and copulas on synthetic and real tabular data through a unified evaluation protocol that reports likelihoods in nats alongside calibration diagnostics.

3. *Implementation.* We document the practical choices that matter in transport-based modelling, including standardisation, variable ordering, and numerical safeguards for cross-term

training, while keeping the introductory chapter focused on the high-level problem.

Two questions motivate the empirical study.

1. *Practical performance.* How do transformation forests perform on synthetic and real tabular data relative to triangular transport maps, copulas, and published normalising-flow baselines? We ask when separable transports reach the performance of high-capacity references and where they fall short once conditional shape depends on context (Rezende and Mohamed, 2015; Dinh *et al.*, 2017; Papamakarios *et al.*, 2021; Hothorn and Zeileis, 2017, 2021).

2. *Trade offs.* What are the trade offs among fit (test NLL), computational efficiency (training wall clock time and per-sample evaluation cost), and calibration diagnostics (probability integral transforms)? We examine where separable transports deliver the best speed-accuracy ratio and where cross-term capacity justifies its added complexity.

1.2 Scope and positioning

We focus on triangular transport maps, transformation random forests, and copulas. Transport maps let us ablate separable versus cross-term capacity under a common objective and shared Jacobian conventions. Transformation forests bring a mature non-parametric estimator whose monotonicity is guaranteed by design and whose induced transport has a clear likelihood interpretation when additive predictors are used. Copulas provide semiparametric and non-parametric dependence models with transparent marginals, serving as interpretable baselines and highlighting elliptical dependence limits in higher dimensions. Normalising flows remain an important reference: we use them for context and draw on published tabular benchmarks because they represent high-capacity autoregressive and coupling architectures widely adopted in density estimation (Rezende and Mohamed, 2015; Dinh *et al.*, 2017; Papamakarios *et al.*, 2021).

1.3 Takeaways and roadmap

A unified transport perspective enables direct comparisons across models with different inductive biases while keeping the introduction free of technical detail. Chapter 2 develops the methodological background on monotone triangular transports, separable versus cross-term structure, and the transport interpretation of transformation forests. Chapter 3 describes the modelling procedures, data handling, and evaluation protocol before presenting empirical results. Chapter 4 synthesises the findings, discusses limitations, and outlines future work.

Chapter 2

Methodological Background

We adopt the tutorial's directionality and notation throughout: data are standardized to $u = (x-\mu)\oslash\sigma$, we learn a monotone triangular map $S : u \mapsto z$ that pushes the standardized target π_U to the reference $\eta = \mathcal{N}(0, I)$, and we evaluate or sample on the original scale via the composition $M(x) = S(T_{\text{std}}(x))$ with the usual affine Jacobian correction when reporting $\log \pi_X(x)$ (subtract $\sum_{k=1}^{K} \log \sigma_k$). All Jacobians and partial derivatives involving S are taken in u-space unless explicitly stated. This chapter develops the minimal change-of-variables machinery, explains why triangular structure is effective, makes precise the contrast between separable and cross-term triangular maps, and shows how transformation random forests (TRTF) instantiate the same transport with an immediate likelihood identity (Rosenblatt, 1952; Knothe, 1957; Hothorn and Zeileis, 2017, 2021; Hothorn *et al.*, 2018; Ramgraber *et al.*, 2025).

2.1 Change of Variables in Standardized Coordinates

The pullback of η by S is the central identity that turns transport into a likelihood,

$$\pi_U(u) = \eta\big(S(u)\big)\,\big|\det \nabla_u S(u)\big|. \tag{2.1}$$

Because we learn S on standardized inputs u, Equation (2.1) is the form we optimize and evaluate. Reporting on the original scale x applies the diagonal Jacobian of T_{std}: $\log \pi_X(x) = \log \pi_U(u) - \sum_{k=1}^{K} \log \sigma_k$ with $u = T_{\text{std}}(x)$. The reference factorizes as $\eta(z) = \prod_{k=1}^{K} \phi(z_k)$, so once S is triangular the log pullback expands as a sum over dimensions. This linear-in-K structure underpins all evaluation and training in later chapters (Ramgraber *et al.*, 2025).

It is useful to record the complementary pushforward statement. If $u \sim \pi_U$ and $z = S(u)$, then $S_{\#}\pi_U = \eta$. The pair $(S^{\sharp}\eta, S_{\#}\pi_U)$ clarifies that the same map both defines the model density in u-space via Equation (2.1) and certifies exact sampling by inversion. In what follows we always compute derivatives of S with respect to u, keep the pullback in standardized coordinates, and only apply the affine correction when moving results back to the original scale.

2.2 Triangular Structure and Its Consequences

A triangular map decomposes component-wise,

$$S(u) = \big(S_1(u_1),\, S_2(u_{1:2}),\, \ldots,\, S_K(u_{1:K})\big), \tag{2.2}$$

and enforces strict monotonicity in the last argument of each component: $\partial_{u_k} S_k(u_{1:k}) > 0$ for all feasible u. The Jacobian $\nabla_u S(u)$ is lower triangular, hence

$$\det \nabla_u S(u) = \prod_{k=1}^{K} \partial_{u_k} S_k(u_{1:k}), \qquad \log \left| \det \nabla_u S(u) \right| = \sum_{k=1}^{K} \log \partial_{u_k} S_k(u_{1:k}). \tag{2.3}$$

Three consequences matter in practice. First, likelihoods are efficient: the log determinant in Equation (2.3) is a sum of one-dimensional terms, so evaluating Equation (2.1) scales in $\mathcal{O}(K)$. Second, invertibility is exact: strict monotonicity implies global bijectivity, and inversion reduces to sequential one-dimensional root finding; compute $u_1 = S_1^{-1}(z_1)$, then $u_2 = S_2^{-1}(z_2; u_1)$, and proceed coordinate by coordinate (Rosenblatt, 1952; Knothe, 1957). Third, triangularity yields transparent conditionals: to sample $u_{m:K} \mid u_{1:m-1} = u^\star_{1:m-1}$, fix the early coordinates at $u^\star$ and invert only the trailing components. Equivalently, the map aligns with the chain-rule factorization $\pi_U(u) = \prod_{k=1}^{K} \pi(u_k \mid u_{1:k-1})$ and exposes each conditional as a one-dimensional transformation (Ramgraber *et al.*, 2025).

Existence and anisotropy deserve a brief remark. For any ordering of the variables there exists a monotone triangular rearrangement that couples π_U and η under weak regularity conditions; this is the Knothe–Rosenblatt rearrangement (Rosenblatt, 1952; Knothe, 1957). Different orderings induce different maps. While all are valid, their sparsity and approximation difficulty may vary with order. When conditional independence makes some inputs irrelevant for a component, omitting those arguments sparsifies $\nabla_u S$, lowers variance, and improves scaling. In high dimension, such sparsity often drives both statistical efficiency and computational cost (Ramgraber *et al.*, 2025).

Finally, we keep terminology consistent. A log density is $\log \hat{\pi}(\cdot)$ at a point. A (test) log likelihood is the dataset average of log densities. Its negative is the negative log likelihood (NLL) in nats; lower is better. These conventions are used uniformly in Chapters **??** and **??**.

2.3 Separable and Cross-Term Triangular Transports

A triangular component is most interpretable once we separate how it depends on the last coordinate u_k versus the context $u_{1:k-1}$. Two parameterizations recur.

Separable triangular maps. A separable component decomposes into a context-dependent shift and a univariate monotone shape,

$$S_k(u_{1:k}) = g_k(u_{1:k-1}) + h_k(u_k), \qquad \partial_{u_k} S_k(u_{1:k}) = h_k'(u_k) > 0, \tag{2.4}$$

so the log Jacobian depends only on u_k:

$$\log \partial_{u_k} S_k(u_{1:k}) = \log h_k'(u_k). \tag{2.5}$$

The context $u_{1:k-1}$ can move the conditional through g_k, but it cannot reshape it. Scale, skewness, tail thickness, and modality remain fixed across contexts because the log-Jacobian contribution depends only on u_k. Separable transports capture nonlinear location shifts across contexts with a context-invariant shape.

Cross-term triangular maps. A cross-term component uses an integrated rectifier,

$$S_k(u_{1:k}) = \int^{u_k} \exp\big(h_k(t, u_{1:k-1})\big)\,\mathrm{d}t + c_k(u_{1:k-1}), \qquad \log \partial_{u_k} S_k(u_{1:k}) = h_k(u_k, u_{1:k-1}), \quad (2.6)$$

so the log Jacobian, and thus the shape of the conditional, depends on the context. This structure captures conditional heteroskedasticity, skew that flips with predictors, or mode splitting that appears only for certain $u_{1:k-1}$.

Two small examples make the contrast concrete. Suppose $U_2 \mid U_1 = u_1$ is normal with mean $m(u_1)$ and variance $\sigma^2(u_1)$. A separable map can absorb $m(u_1)$ through g_2, but cannot make the slope $\partial_{u_2} S_2$ depend on u_1, so it cannot represent the variance change $\sigma^2(u_1)$. A cross-term map sets $\log \partial_{u_2} S_2 = h_2(u_2, u_1)$ and can encode variance inflation or contraction with u_1. As a second example, let $U_2 \mid U_1$ be unimodal for some u_1 and bimodal for others. Separable maps preserve modality across contexts; cross terms allow the derivative to bend with u_1 and can introduce or remove modes as the context varies. Cross-term training requires careful numerical safeguards; details follow in Chapter **??**.

Implementation notes. The integrated rectifier guarantees positivity of $\partial_{u_k} S_k$ and supports rich context dependence, but typically requires one-dimensional quadrature and regularization to remain stable. Separable maps often admit linear-in-coefficient parameterizations with convex subproblems for the monotone part; they are fast and robust, but limited to context-invariant conditional shape (Ramgraber *et al.*, 2025).

2.4 Transformation Forests as Triangular Transport

Transformation models posit a strictly increasing transformation $h(y \mid w)$ such that $\Phi\big(h(Y \mid W)\big)$ is standard, letting predictors act through h while preserving monotonicity (Hothorn *et al.*, 2018). A transformation random forest aggregates local transformation models over an adaptive partition of the predictor space to produce a strictly monotone conditional CDF $\widehat{F}_k(\cdot \mid u_{1:k-1})$ for each coordinate u_k given $u_{1:k-1}$ (Hothorn and Zeileis, 2017, 2021). This induces a triangular transport component via the probability integral transform,

$$S_k(u_{1:k}) = \Phi^{-1}\big(\widehat{F}_k(u_k \mid u_{1:k-1})\big). \quad (2.7)$$

Per-component likelihood identity. Differentiating $\Phi\big(S_k(u_{1:k})\big) = \widehat{F}_k(u_k \mid u_{1:k-1})$ in u_k gives

$$\widehat{\pi}_k(u_k \mid u_{1:k-1}) = \phi\big(S_k(u_{1:k})\big)\,\partial_{u_k} S_k(u_{1:k}), \quad (2.8)$$

showing that the TRTF conditional density equals the pullback factor one would obtain by parameterizing S directly. Summing over k yields the joint log likelihood from Equation (2.1) with the triangular log determinant in Equation (2.3). TRTF therefore implements the same transport likelihood once standardized (Hothorn and Zeileis, 2017, 2021; Ramgraber *et al.*, 2025).

Additive predictor implies separability. Under the common additive predictor implementation (TRTF-AP),

$$\widehat{F}_k(u_k \mid u_{1:k-1}) = \Phi\big(h_k(u_k) + g_k(u_{1:k-1})\big), \quad (2.9)$$

so

$$S_k(u_{1:k}) = h_k(u_k) + g_k(u_{1:k-1}), \qquad \partial_{u_k} S_k(u_{1:k}) = h_k'(u_k), \tag{2.10}$$

and the induced transport is separable: $\log \partial_{u_k} S_k$ depends only on u_k. Monotonicity in u_k is ensured by construction, and the Jacobian term carries no context dependence. TRTF-AP therefore excels when conditional shapes are stable and context acts primarily through location; it is structurally limited when shape varies with predictors (Hothorn and Zeileis, 2017, 2021; Hothorn *et al.*, 2018).

A short note on reporting closes the loop. TRTF evaluates conditional densities in standardized coordinates. Reported log densities on the original scale apply the affine correction $-\sum_{k=1}^{K} \log \sigma_k$ that accompanies T_{std}, matching the convention set at the beginning of this chapter. With this alignment, comparisons between TRTF, direct triangular maps, and copulas are numerically coherent.

2.5 Takeaways

We learn $S : u \mapsto z$ in standardized coordinates and evaluate $\log \pi_X$ via a simple affine correction. Triangularity turns the log determinant into a sum of one-dimensional terms, guarantees exact inversion, and exposes conditionals through back substitution; these properties make likelihoods linear time in dimension and conditional simulation straightforward (Rosenblatt, 1952; Knothe, 1957; Ramgraber *et al.*, 2025). Separable transports move locations but keep shape fixed across contexts; cross-term transports allow context-dependent shape and capture heteroskedastic or multimodal conditionals at the cost of more delicate optimization. TRTF provides a nonparametric route to the same triangular transport with a per-component likelihood identity and the important specialization that TRTF-AP is separable by construction (Hothorn and Zeileis, 2017, 2021; Hothorn *et al.*, 2018). These foundations justify the objectives and implementations developed in Chapter **??** and frame the empirical comparisons that follow.

Chapter 3

Data Analysis and Validation

This chapter turns the commitments of Chapters 1 and 2 into a practical modeling programme. 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.

3.1 Setup: Standardization, Map Direction, and Reporting

Observations on the original scale satisfy $x \in \mathbb{R}^K$. We standardize using training-split statistics

$$u = T_{\text{std}}(x) = (x - \mu) \oslash \sigma, \qquad \sigma_k > 0, \tag{3.1}$$

learn a single monotone triangular map in this standardized space,

$$S : u \longmapsto z, \qquad z \sim \eta = \mathcal{N}(0, I), \tag{3.2}$$

and sample via $S^{-1} : z \mapsto u \mapsto x$. All Jacobians and partial derivatives that involve S are taken with respect to u unless explicitly stated. The pullback identity (2.1) remains the objective we optimize.

Reported log-densities on the original scale apply the diagonal affine correction

$$\log \pi_X(x) = \log \pi_U\big(T_{\text{std}}(x)\big) - \sum_{k=1}^{K} \log \sigma_k, \tag{3.3}$$

so every method produces pointwise values $\log \hat{\pi}_X(x)$ that are directly comparable. We reserve the term *log-density* for the value at a single point, *(test) log-likelihood* for the dataset average, and *negative log-likelihood (NLL)* for its negative; NLL is always reported in nats.

The choice of direction $S : u \mapsto z$ keeps the change-of-variables identity separable, giving per-sample complexity $\mathcal{O}(K)$ through the triangular determinant in Equation (2.3). The same structure yields exact inversion—solve K one-dimensional monotone equations in sequence—and transparent conditionals by fixing $u_{1:m-1}$ and inverting the trailing components (Rosenblatt, 1952; Knothe, 1957). Because every method uses the same direction, we avoid mixing objectives

or Jacobian conventions and can incorporate copulas, whose natural domain is x-space, via the standardization wrapper.

3.2 Triangular Transport Maps

With standardized coordinates fixed, we describe the transports we fit directly.

3.2.1 Separable and Cross-Term Components

A triangular map factorizes as in Equation (2.2) with strictly positive diagonal derivatives. Two parameterizations anchor the expressiveness spectrum.

Separable components. These decompose into a context shift and a univariate monotone shape,

$$S_k(u_{1:k}) = g_k(u_{1:k-1}) + h_k(u_k), \qquad \log \partial_{u_k} S_k(u_{1:k}) = \log h_k'(u_k), \tag{3.4}$$

so the log-Jacobian term depends only on u_k. The context can shift location through g_k but cannot reshape the conditional: scale, skewness, tail thickness, and modality stay invariant across contexts.

Cross-term components. To capture context-dependent shape we use the integrated-rectifier construction

$$S_k(u_{1:k}) = \int^{u_k} \exp\big(h_k(t, u_{1:k-1})\big)\,\mathrm{d}t + c_k(u_{1:k-1}), \qquad \log \partial_{u_k} S_k(u_{1:k}) = h_k(u_k, u_{1:k-1}), \tag{3.5}$$

so the log-Jacobian depends on both u_k and the context. This structure captures conditional heteroskedasticity, skew changes, and context-specific multimodality at the cost of a more delicate numerical treatment.

Two examples highlight the contrast. If $U_2 \mid U_1 = u_1$ is Gaussian with mean $m(u_1)$ and variance $\sigma^2(u_1)$, a separable map can absorb $m(u_1)$ through g_2 but cannot reflect the variance change; a cross-term map adjusts $h_2(u_2, u_1)$ to encode variance inflation or contraction. If $U_2 \mid U_1$ switches between unimodal and bimodal shapes across contexts, separable maps preserve modality, whereas cross-term components can introduce or remove modes as u_1 varies.

3.2.2 Monotone Parameterizations

We rely on two monotonicity mechanisms that mirror the structures above.

Linear-in-coefficients (LIC) separable maps. We expand h_k in monotone one-dimensional bases—identity, integrated sigmoids, softplus-like edge terms, integrated radial basis functions—with nonnegative coefficients. This ensures $h_k'(u_k) \geq 0$ by construction and yields fast, stable inner loops: the monotone part often reduces to a convex subproblem, while g_k updates via least squares.

Integrated-rectifier cross-term maps. We parameterize $h_k(u_k, u_{1:k-1})$ directly, apply a rectifier such as the exponential to enforce positivity, and integrate over u_k. The rectifier guarantees strictly positive derivatives, the integral maintains global monotonicity in the last coordinate, and the context arguments allow rich dependence. The expressiveness is necessary for heteroskedastic, skewed, or multimodal conditionals, but it makes optimization numerically sensitive.

3.2.3 Basis Choices and Tail Behaviour

Both parameterizations combine global and local bases with explicit tail control:

- **Hermite functions.** Hermite polynomials align naturally with Gaussian references. We favour Hermite functions—polynomials multiplied by Gaussian weights—so that higher-order terms taper in the tails, while retaining linear terms unweighted to keep S_k linear as $|u_k|$ grows. Tail linearization stabilizes inversion by preventing vanishing or exploding derivatives.

- **Localized bases.** Integrated sigmoids, softplus edges, and integrated radial basis functions capture local features. In separable maps they assemble monotone $h_k(u_k)$; in cross-term maps they populate $h_k(u_k, u_{1:k-1})$ to bend the log-derivative with context. We place centres at empirical quantiles and set widths via nearest-neighbour distances for well-spread coverage without per-fit location parameters.

3.2.4 Ordering, Sparsity, and Objectives

Triangular structure is anisotropic: the variable ordering matters. The Knothe–Rosenblatt rearrangement guarantees a monotone triangular coupling for any order, but the sparsity and approximation difficulty of a finite-basis parameterization can change dramatically (Rosenblatt, 1952; Knothe, 1957). We use the natural data ordering in primary results and drop arguments from S_k whenever conditional independence is plausible. The Jacobian becomes sparser, evaluation cheaper, and variance lower in small-sample regimes.

With $z \sim \mathcal{N}(0,1)$ i.i.d., $\log \eta(S(u)) = -\frac{1}{2}\sum_k S_k(u_{1:k})^2 - \frac{K}{2}\log(2\pi)$. Optimizing the pullback likelihood therefore amounts to minimising the separable objective

$$\sum_{k=1}^{K} \left[\tfrac{1}{2} S_k(u_{1:k})^2 - \log \partial_{u_k} S_k(u_{1:k})\right], \tag{3.6}$$

which we treat as the workhorse loss for direct TTM training: a quadratic push-to-Gaussian term and a log-barrier that forbids vanishing derivatives. Equation (3.6) will reappear as an identity for TRTF.

3.2.5 Safeguards for Cross-Term Training

The integrated-rectifier parameterization is expressive, but three safeguards are critical for stability, especially at large K or with heavy-tailed data.

1. **Quadrature.** We evaluate the integral in Equation (3.5) via Gauss–Legendre quadrature with a dataset-tuned node count.

2. **Log-derivative clipping.** We clip h_k to a bounded interval $[-H, H]$ inside the rectifier, capping $\partial_{u_k} S_k$ between $\exp(-H)$ and $\exp(H)$. This prevents overflow, stabilizes the log-determinant, and avoids near-flat tails that hinder inversion.

3. **Regularization.** We apply mild ridge penalties to coefficients that parameterize h_k (and when useful to c_k). Regularization stabilizes identification in sparse regions and interacts well with clipping: parameters that push h_k outside $[-H, H]$ meet both a hard cap and a quadratic penalty. When combined with tail linearization, these levers make cross-term optimisation behave like a well-tempered extension of the separable case.

Once trained, inversion is sequential and exact: solve $u_1 = S_1^{-1}(z_1)$, then $u_2 = S_2^{-1}(z_2; u_1)$, and so forth. Each step is a monotone root-find with robust bracketing thanks to tail linearization. Conditionals come for free by fixing $u_{1:m-1}$, drawing $z_m, \ldots, z_K \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and inverting the trailing block. Likelihood evaluation, inversion, and conditional sampling all reduce to sums and one-dimensional operations, so per-sample complexity remains linear in K.

3.3 Transformation Random Forests as Transport

Transformation models posit a strictly increasing transformation $h(y \mid w)$ such that $\Phi\big(h(Y \mid W)\big)$ is standard (Hothorn *et al.*, 2018). A transformation random forest aggregates local transformation models over adaptive partitions to produce strictly monotone conditional CDFs $\widehat{F}_k(\cdot \mid u_{1:k-1})$ (Hothorn and Zeileis, 2017, 2021). The induced triangular transport is given by Equation (2.7), and differentiating yields the likelihood identity (2.8). Summing over k recovers the same pull-back objective as Equation (3.6) once standardized.

Under the ubiquitous additive predictor implementation,

$$\widehat{F}_k(u_k \mid u_{1:k-1}) = \Phi\big(h_k(u_k) + g_k(u_{1:k-1})\big), \tag{3.7}$$

so $S_k(u_{1:k}) = h_k(u_k) + g_k(u_{1:k-1})$ and $\partial_{u_k} S_k(u_{1:k}) = h_k'(u_k)$. The induced transport is separable; context shifts location but cannot change shape. Monotonicity in u_k is guaranteed by construction, forests mitigate variance through aggregation, and the same back-substitution used for TTM provides inversion. Wherever conditional variance, skewness, or modality depends on the predictors, TRTF-AP is structurally under-specified.

The construction does not preclude non-additive predictors that would reintroduce cross-term capacity, but the standard implementation—used here for robustness and interpretability—adopts the additive predictor. This makes the link to separable TTM explicit and frames the central empirical question: when is separability sufficient, and when does cross-term capacity pay off?

3.4 Copula Baselines

Copulas decouple marginals from dependence via Sklar's theorem (Sklar, 1959). They offer interpretable baselines with transparent marginal modelling. We employ two recipes.

- **Semiparametric Gaussian copula.** We compute rank-based pseudo-observations $\hat{v}_{ik}$, transform them via $z_{ik} = \Phi^{-1}(\hat{v}_{ik})$, estimate a correlation matrix $\widehat{\Sigma}$ (with regularisation if

needed), and evaluate the Gaussian copula density using the multivariate normal pdf on z. The joint density is $\hat{\pi}(x) = c_{\text{Gauss}}\big(\hat{v}(x); \widehat{\Sigma}\big) \prod_k \hat{\pi}_k(x_k)$. This scales gracefully to high K but enforces elliptical dependence.

- **Low-dimensional nonparametric copula.** For small K we fit a kernel density estimator on the probit-transformed pseudo-observations and map back with the appropriate Jacobian. This avoids parametric dependence assumptions but suffers from the curse of dimensionality, so we restrict it to $K \leq 3$.

These copulas act as interpretable dependence baselines and highlight when localized, context-dependent changes are essential.

3.5 Evaluation Protocol

We evaluate all models under a common rubric.

3.5.1 Log-Likelihoods and NLL

For any fitted model $\hat{\pi}$, the log-density at a point is $\log \hat{\pi}_X(x)$. The test log-likelihood averages this quantity over the test split, and the NLL is its negative. For triangular models we exploit the decomposition

$$\log \hat{\pi}_U(u) = \sum_{k=1}^{K} \Big[\log \phi\big(S_k(u_{1:k})\big) + \log \partial_{u_k} S_k(u_{1:k}) \Big], \tag{3.8}$$

with $u = T_{\text{std}}(x)$, before applying the correction (3.3). The additive structure lets us report per-dimension conditional NLLs

$$\text{NLL}_k \approx -\frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \log \hat{\pi}\big(x_{ik} \mid x_{i,1:k-1}\big), \tag{3.9}$$

which sum to the joint NLL and identify difficult conditionals. Copulas do not admit a unique triangular factorisation, so we report only their joint NLL.

3.5.2 Calibration via PIT

Calibration asks whether probabilities are correct, especially conditionals. For a well-calibrated triangular model, the conditional PIT values

$$\widehat{V}_{ik} = \widehat{F}_k\big(u_{ik} \mid u_{i,1:k-1}\big) \tag{3.10}$$

should be i.i.d. $\text{Unif}(0,1)$ over the test set. We visualise $\{\widehat{V}_{ik}\}$ with PIT histograms per k and optionally report summary statistics such as the Kolmogorov–Smirnov distance to uniformity. For copulas we assess calibration of the marginals and low-dimensional slices where dependence is most transparent. Systematic departures (U-shaped or inverse-U PIT) flag under- or over-dispersion; for separable transports they often indicate that cross-term capacity would help.

3.5.3 Compute Metrics and Memory

We record wall-clock training time and per-sample evaluation time on the test set. For TTMs, both scale linearly in K and approximately linearly in the number of basis functions; cross-term maps incur a small constant-factor overhead from one-dimensional quadrature. TRTF training time scales with the number and depth of trees per conditional, while prediction retains linear scaling after aggregation. Copula training is dominated by correlation estimation or KDE fitting, with fast evaluation. Where relevant we also log memory footprints, as some implementations trade RAM for speed via cached basis evaluations or forest leaf summaries. All methods use the same training-only standardisation parameters (μ, σ); seeds are fixed across splits; and we report standard errors over multiple runs to quantify stochastic variability.

3.5.4 Defaults and External Baselines

To keep the main text readable we adopt consolidated defaults. For cross-term TTMs we use Gauss–Legendre quadrature, clip log-derivatives to $[-H, H]$, and apply mild L_2 penalties (with optional L_1 on context shifts) tuned on validation. We use the data's natural variable ordering for headline results and examine robustness across a few alternatives elsewhere. Normalizing flows provide contextual baselines from the literature: they compose invertible layers with permutations or autoregressive sublayers (Rezende and Mohamed, 2015; Dinh *et al.*, 2017; Papamakarios *et al.*, 2021), achieve strong likelihoods, but do not enforce strict triangular structure; we include published flow results on MINIBOONE only as external reference lines.

3.6 Data and Preprocessing

We evaluate the models on synthetic datasets (Half-Moon and a four-dimensional conditional generator) and on the real-world MINIBOONE dataset.

MINIBOONE. To ensure comparability with published normalising-flow benchmarks we follow the preprocessing protocol of Papamakarios *et al.* (2017). The steps are: removing 11 outliers with the value -1000; dropping seven features with extreme concentration on a single value; yielding a final dimensionality of $K = 43$; using the fixed train/validation/test splits provided in the benchmark (Appendix D/E of Papamakarios *et al.* (2017)); applying train-only standardisation as described in Section 3.1; and refraining from any additional pruning of highly correlated features, as that would break comparability of log-likelihoods across studies.

Synthetic data. The synthetic experiments use analytically specified data-generating processes. Details for the four-dimensional conditional generator appear in Appendix A.

3.7 Metrics and Baselines

Goodness of fit. The primary evaluation metric is the average test negative log-likelihood (NLL) in nats (lower is better).

Conditional NLL. For triangular models (TTM, TRTF) we exploit the decomposition implied by Equation (3.8) to report per-dimension negative log-likelihoods,

$$\mathrm{NLL}_k \approx -\frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \log \hat{\pi}\big(x_{ik} \mid x_{i,1:k-1}\big), \tag{3.11}$$

which sum to the joint NLL and identify difficult conditionals. For non-triangular models such as copulas the decomposition is not unique, so we report only the joint NLL.

Calibration. We assess calibration using probability integral transform (PIT) histograms of the conditional CDF values $\widehat{V}_{ik}$ in Equation (3.10). Uniform PIT indicates well-calibrated conditionals; U- or inverted-U shapes flag under- or over-dispersion. For copulas we examine marginal and low-dimensional calibration where the dependence structure is most transparent.

Baselines. The experiments include two reference points: an independent baseline (product of estimated marginals, e.g. TTM-D) and, for simulations only, oracle models based on the true joint density or true marginals.

3.8 Robustness and Sensitivity

All experiments use deterministic seeds; we average results over multiple runs and report standard errors. We gauge practical significance via overlap of $\pm 2\,\mathrm{SE}$ intervals. Variable ordering matters for triangular estimators—the Knothe–Rosenblatt map exists for any order but finite-basis approximations can vary in difficulty (Ramgraber *et al.*, 2025). We report results for the natural data ordering and note opportunities to explore heuristics across alternative orders.

3.9 Empirical Results

This section reports empirical findings for the three modelling families within the common transport frame established above. All models learn a monotone triangular map $S : u \to z$ on train-only standardized inputs, evaluate Jacobians in u-space, and report log-densities on the original scale with the fixed affine correction. We summarize results by test negative log-likelihood (NLL, nats; lower is better) and, for triangular models, by per-dimension conditional NLLs that sum to the joint. Where informative, we discuss calibration through probability integral transform (PIT) diagnostics and comment on computation. External flow results on MINIBOONE are included only as context from the literature.

3.9.1 Synthetic Data

Half-Moon ($K = 2$)

Setup. The two interleaving half-circles with additive Gaussian noise produce a bimodal joint with strong curvature. We fit TTM-Marginal, TTM-Separable (TTM-S), TTM-Cross-term (TTM-X), TRTF under an additive predictor (TRTF-AP), and a low-dimensional nonparametric copula with probit transform and KDE dependence. Two references provide context: a true

joint likelihood based on the known mixture and a class-conditional oracle $p(x \mid y)$ using the true component label.

Results. Table 3.1 reports joint and conditional NLLs. TTM-X markedly improves over separable and marginal transports, consistent with the need for context-dependent shape. TRTF-AP behaves like a separable transport by construction and underperforms when conditional shape depends on the context, matching the theory in Chapter 2. The nonparametric copula performs strongly in two dimensions, as expected for a flexible bivariate dependence model with well-calibrated marginals.

Table 3.1: Half-Moon test negative log-likelihoods (nats). Lower is better; ± indicates standard error.

Model	Mean joint NLL	Conditional NLL 1	Conditional NLL 2
Copula-NP	0.87 ± 0.16	0.76	0.11
TTM-X	1.22 ± 0.20	0.92	0.29
TRTF-AP	1.83 ± 0.14	1.25	0.57
TTM-S	1.92 ± 0.14	1.29	0.64
TTM-Marginal	2.04 ± 0.12	1.29	0.75
True marginals	1.37 ± 0.11	0.69	0.69
True joint	0.70 ± 0.12	0.35	0.35

Four-Dimensional Autoregressive Generator

Setup. The synthetic four-dimensional task combines heteroskedastic, skewed, and conditional multimodal components over a range of sample sizes. We compare the same suite of models as above.

Results. At very small n the flexible cross-term TTM-X is variance limited, and simpler models (TTM-S, TRTF-AP) can be competitive. As n grows, TTM-X consistently approaches the true joint likelihood, confirming the benefit of cross-term capacity. Conditional NLL decompositions highlight which dimensions drive residual error for separable transports; Table 3.2 illustrates the small-sample regime.

Table 3.2: Four-dimensional autoregressive generator ($n = 50$): conditional and joint NLLs (nats). Values are means with standard errors.

Dim	Distribution	True marg.	True joint	TRTF-AP	TTM-Marginal	TTM-S	TTM
1	Normal	1.46 ± 0.26	1.41 ± 0.29	1.48 ± 0.24	1.49 ± 0.20	1.46 ± 0.26	1.46 ± 0
2	Exponential	1.55 ± 0.46	1.38 ± 0.65	2.54 ± 0.75	3.30 ± 0.01	1.75 ± 0.70	2.58 ± 0
3	Beta	-0.46 ± 0.63	-0.63 ± 1.00	-0.14 ± 0.34	0.40 ± 0.17	0.28 ± 0.70	0.39 ± 0
4	Gamma	2.21 ± 1.11	2.07 ± 0.80	2.22 ± 1.08	2.78 ± 0.77	2.97 ± 1.45	3.00 ± 1
K	Sum (joint)	4.75 ± 1.10	4.23 ± 1.03	6.10 ± 1.61	7.97 ± 0.94	6.47 ± 1.92	7.43 ± 1

3.9.2 Real Data: MINIBOONE

Setup. We train all models on the standardized MINIBOONE tabular dataset ($K = 43$) with the protocol from Papamakarios *et al.* (2017). For comparison we include published NLL results from high-capacity normalising flows.

Results. Table 3.3 summarises average test log-likelihoods reported in nats (higher is better). TRTF-AP improves markedly over an independent Gaussian baseline, yet it trails deep flows reported in the literature, which is consistent with the separable Jacobian constraint.

Table 3.3: MINIBOONE test log-likelihoods (nats, higher is better). Literature numbers follow Papamakarios *et al.* (2017).

Model	Test log-likelihood
Gaussian independent baseline	-37.24 ± 1.07
TRTF-AP (this study)	-29.88 ± 0.02
MADE with auxiliary conditioning	-15.59 ± 0.50
Real NVP (five layers)	-13.55 ± 0.49
MAF (five layers)	-11.75 ± 0.44
MAF mixture of Gaussians (five layers)	-11.68 ± 0.44

Interpretation. TRTF-AP captures meaningful dependence relative to the independent baseline, but deep flows retain a substantial advantage on this dataset. The gap aligns with the additional flexibility of autoregressive and coupling architectures documented by Papamakarios *et al.* (2017) and Dinh *et al.* (2017). Conditional PIT diagnostics show that TRTF-AP delivers calibrated marginals by design and partially calibrated conditionals; cross-term transports narrow the remaining deviations at additional computational cost.

Compute. Triangular models exhibit linear-time evaluation in K. Cross-term TTM incurs modest overhead from one-dimensional quadrature and clipping safeguards. TRTF training scales with the number and depth of trees per conditional; evaluation is an ensemble average. Copula training is dominated by correlation estimation or KDE fitting, with fast evaluation thereafter. We additionally report memory footprints where caching affects runtime.

Summary. Across synthetic and real datasets, allowing context-dependent shape via cross-term derivatives consistently improves fit and calibration. Separable transports—including TRTF-AP—remain attractive when structure is near-separable or when computational stability is paramount. Copulas offer interpretable baselines but lag in higher dimensions where elliptical or low-dimensional assumptions are violated. The next chapter interprets these findings and discusses practical implications.

3.10 Takeaways

The methodological backbone is deliberately uniform: learn a single monotone triangular map $S : u \mapsto z$ in standardized coordinates, keep all Jacobians and derivatives in u-space, and report log-densities on x with the fixed affine correction (3.3). Within that frame, separable transports

shift location but keep shape fixed across contexts; cross-term transports allow context-dependent shape via quadrature, clipping, and regularisation; TRTF induces the same triangular likelihood via $S_k = \Phi^{-1}(\widehat{F}_k)$, with the additive predictor implying separability exactly; and copulas serve as interpretable dependence baselines. We judge models by NLL (nats), PIT calibration, and compute, enabling a clean assessment of when separability is right-sized, when cross-term capacity pays off, and how TRTF and copulas position themselves relative to direct triangular transports and high-capacity flows (Rosenblatt, 1952; Knothe, 1957; Hothorn and Zeileis, 2017, 2021; Hothorn *et al.*, 2018; Ramgraber *et al.*, 2025; Sklar, 1959; Rezende and Mohamed, 2015; Dinh *et al.*, 2017; Papamakarios *et al.*, 2021).

Chapter 4

Interpretation and Conclusion

This chapter synthesizes the empirical evidence gathered in Chapter 3, highlights methodological insights, and outlines practical guidance and future work.

4.1 Summary of Findings and Trade-offs

4.1.1 Synthetic benchmarks

For the Half-Moon dataset ($K = 2$), cross-term triangular transport maps (TTM-X) outperform separable transports and transformation forests. This performance confirms the need for context-dependent shape when slices through the data change modality along the curve (Table 3.1). A low-dimensional nonparametric copula performs competitively in two dimensions, as expected for a flexible bivariate dependence model with well-calibrated marginals. Simpler models (TTM-S, TRTF-AP) remain competitive at small sample sizes because their variance is lower. The four-dimensional autoregressive generator highlights this variance effect (Table 3.2). TTM-X consistently approaches the true joint likelihood as the sample size grows (Table 3.2). Conditional NLL decompositions highlight which dimensions drive residual error for separable transports.

4.1.2 MINIBOONE

The additive-predictor transformation forest (TRTF-AP) improves over an independent baseline on the real-world MINIBOONE dataset ($K = 43$). The estimator remains behind state-of-the-art normalizing flows, consistent with the separable Jacobian constraint (Table 3.3). Cross-term TTM variants reduce the remaining gap. These variants require additional optimization safeguards such as quadrature, clipping, and regularization. Probability integral transform (PIT) diagnostics show that models with context-dependent shape deliver more uniform conditional PITs than separable alternatives.

Cross-term parameterizations improve fit and calibration across datasets by allowing the log-derivative to depend on context. Separable transports, including TRTF-AP, remain attractive when structure is approximately separable or when computational stability is paramount. Copulas serve as interpretable dependence baselines but lag in higher dimensions where elliptical or low-dimensional assumptions are violated.

4.2 Methodological Insights

Transformation forests and triangular transport are tightly linked in this modeling framework. With additive predictors, TRTF induces separable maps that guarantee monotonicity and interpretability but cannot capture heteroskedasticity or multimodal conditionals. Cross-term triangular maps restore context-dependent shape through log-derivatives and require safeguards such as quadrature, clipping, mild ridge penalties, and tail linearization. Copulas decouple marginals from dependence and are most useful when elliptical structure is adequate.

4.3 Practical Guidance

First, use separable transports or TRTF-AP when conditional shapes are approximately invariant across contexts, or when interpretability and stability take priority. Then, deploy cross-term maps when calibration or likelihood diagnostics indicate context-dependent variance, skewness, or multimodality, and enable quadrature, clipping, and mild regularization. Moreover, treat copulas as interpretable baselines and diagnostic tools for elliptical or low-dimensional dependence. Finally, keep train-only standardization, consistent Jacobian conventions, and the diagonal affine correction so reported log-densities remain comparable.

4.4 Limitations and Future Work

Key limitations center on ordering sensitivity, richer parameterizations, extended TRTF variants, and high-dimensional copulas. Triangular transports are anisotropic; data-driven ordering heuristics or permutation-invariant architectures may reduce variance and improve scalability. Richer neural parameterizations could bridge the gap to modern flows while preserving exact likelihoods. Non-additive predictors in transformation forests could restore cross-term capacity while retaining ensemble robustness. Copulas with context-aware dependence may offer interpretable yet flexible alternatives in higher dimensions.

4.5 Concluding Remarks

This thesis established a unified transport-based framework for comparing triangular transport maps, transformation forests, and copulas on synthetic and real tabular data. Recognizing when separability suffices and when cross-term capacity is necessary helps practitioners balance interpretability, computational cost, and statistical performance. Continued advances in transport parameterizations, ensemble methods, and hybrid copulas promise further gains in modeling complex conditional densities.

Chapter 5

References

Bibliography

Dinh, L., Krueger, D., and Bengio, Y. (2017). Density estimation using real NVP. In *International Conference on Learning Representations*. 2, 12, 15, 16

Hothorn, T., Kneib, T., and B"uhlmann, P. (2018). Conditional transformation models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**, 955–980. 1, 3, 5, 6, 10, 16

Hothorn, T. and Zeileis, A. (2017). Transformation forests. *Machine Learning*, **106**, 1469–1481. 1, 2, 3, 5, 6, 10, 16

Hothorn, T. and Zeileis, A. (2021). Transformation forests: A framework for parametric, nonparametric, and semiparametric regression and distributional modeling. Working paper. 1, 2, 3, 5, 6, 10, 16

Knothe, H. (1957). Contributions to the theory of convex bodies. *Mathematische Zeitschrift*, **66**, 199–210. 3, 4, 6, 7, 9, 16

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, **22**, 1–64. 2, 12, 16

Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, volume 30. 12, 15

Ramgraber, M.*et al.* (2025). A friendly introduction to triangular transport. Preprint. 3, 4, 5, 6, 13, 16

Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, 1530–1538. 2, 12, 16

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, **23**, 470–472. 3, 4, 6, 7, 9, 16

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, **8**, 229–231. 10, 16

Appendix A

Appendix

Maybe some R code here, probably a `sessionInfo()`